University of Zurich[UZH]

Speaker Idiosyncratic Intensity Variability in the Speech Signal

Thesis (cumulative thesis)
presented to the Faculty of Arts and Social Sciences
of the University of Zurich
for the degree of Doctor of Philosophy

by Lei He

Accepted in the spring semester 2016
on the recommendation of the doctoral committee:
Prof. Dr. Volker Dellwo (main supervisor)
Prof. Dr. Martin Meyer

Zurich, 2017

Copyright notice

Acknowledgements

The completion of this work was made possible through the guidance, help and support from a various number of friends, colleagues and family. First and foremost, I would like to extend my heartfelt thanks to Prof. Dr. Volker Dellwo for giving me the opportunity to work at the University of Zurich (UZH), and his constant guidance, support and inspirations. Many thanks go to Prof. Dr. Martin Meyer whose advice to situate my current and future research in a wider academic context is invaluable. I am also thankful to Prof. Dr. Stephan Schmid for his helpful comments and suggestions throughout my PhD research.

Working in the phonetics group at UZH is a privilege. I had many enlightening and enjoyable times talking and discussing with my fellow colleagues and friends: Kostis Dimos, Daniel Friedrichs, Thayabaran Kathiresan, Elisa Pellegrino and Sandra Schwab. Particular thanks go to Sascha Völlmin for his various helps in the Doktoratsprogramm Linguistik (DPL) at UZH, and Marie-José Kolly and Adrian Leemann for their work in constructing the TEVOID corpus which I used extensively for my research.

Special thanks go to Prof. Dr. Rushen Shi. During her sabbatical stay in Zurich, I received much helpful advice from her on maintaining a sustained research career. We had many enjoyable conversations about our home country as well.

Prof. Dr. Hongyun Wu at Renmin University of China has been a constant supporter of my academic career ever since I was doing my master's study with her. Her passion for research and knowledge inspired me to embark on an academic journey.

It was impossible to complete my PhD without the support of my family. To my parents and parents-in-law, thank you for making everything so easy for me so that I can concentrate on my studies. To my loving wife, Yu Zhang, thank you for all the sacrifices and compromises during the past years while we were geographically separated. Thank you for having faith in me! To my grandma, you had always supported and believed in me ever since I was little. You left us a few days before my defense, but you are never really gone in my heart.

Last by not least, I thank the Gebert Rüf Stiftung for supporting the project I was working in, and the Swiss Academy of Humanities and Social Sciences (SAGW) as well as the DPL at UZH for the numerous financial supports for travel expenses.
■

Table of contents

Abstract

This cumulative thesis addresses how speaker individual differences are reflected in the intensity variability in the speech signal. In **Study I**, characteristics of intensity variability (average or peak) between syllables were measured. The results indicated significant effects of the speakers in all intensity measures. **Study II** compares the speaker recognition strengths based on suprasegmental duration and intensity variability in the speech signal using artificial neural networks. The results indicated that both intensity and combined metrics significantly outperformed the duration measures. **Study III** examines the role of syllabic intensity characteristics in between-speaker rhythmic variability. It was found that the intensity measures varied significantly between speakers. A semiautomatic speaker recognition based on duration and intensity measures using multinomial logistic regression and feedforward neural networks was carried out. Results showed that intensity measures contained stronger speaker specific information compared to measures based on durational variability of phonetic intervals. In addition, effects of the recognition algorithms and data normalization procedures were discovered. In **Study IV**, intensity contours of speech signals were sub-divided into positive and negative dynamics. Mean, standard deviation, and sequential variability were measured for both dynamics in each sentence. Analyses showed that measures of both dynamics were separately classified and between-speaker variability was largely explained by measures of negative dynamics. This suggests that parts of the signal where intensity decreases from syllable peaks are more speaker-specific. Idiosyncratic articulation may explain such results.

■

Zusammenfassung

Dieser kumulativen Dissertation liegt die Fragestellung zu Grunde, in welcher Weise sprecherspezifische Unterschiede in der Variabilität des akustischen Sprachsignals reflektiert sind. In **Studie I**, Merkmale von Intensitätsvariabilität (Durchschnitts- oder Spitzenwerte) zwischen Silben wurden. Die Ergebnisse zeigten signifikante Effekte für die Sprecher für alle untersuchten Merkmale. **Studie II** vergleicht die Zuverlässigkeit der Sprechererkennung basierend auf suprasegmentalen Zeitmassen bzw. Dauern und der Intensitätsvariabilität im Sprachsignal mit Hilfe künstlicher neuronaler Netze. Die Ergebnisse haben gezeigt, dass Intensitätsmasse und kombinierte Masse signifikant bessere Resultate ermöglichen. **Studie III** untersucht die Rolle silbischer Intensitätscharakteristika für die rhythmische Variabilität zwischen Sprechern (Inter-Sprecher-Variabilität). Eine halbautomatische Sprechererkennung basierend auf Zeitmassen und Intensitätskorrelaten wurde mit Hilfe multinominaler logischer Regression und neuronaler Feedforward-Netze. Die Ergebnisse haben gezeigt, dass Intensitätskorrelate robustere sprecherspezifische Informationen transportieren als Korrelate, die auf der Variabilität von Zeitmassen phonetischer Intervalle beruhen. Des Weiteren wurden Effekte des Erkennungsalgorithmus und der Daten-Normalisierungs-Prozeduren gefunden. In **Studie IV** wurden Intensitätsverläufe im Sprachsignal in positive und negative Dynamiken unterteilt. Für beide Dynamiken wurden in jedem Satz Durchschnitt, Standardabweichung und sequentielle Variabilität gemessen. Die Analysen haben gezeigt das Masse beider Dynamiken separat klassifiziert wurden und die Inter-Sprecher-Variabilität grösstenteils durch Masse negativer Dynamiken zu erklären war. Dies lässt vermuten, dass die Teile des Signals sprecherspezifische sind, in denen die Intensität von Silbenspitzen abfällt. Idiosynkratische Artikulation dürfte derartige Ergebnisse erklären. (*Translation from the English Abstract by Daniel Friedrichs*).

■

Synopsis

This cumulative thesis consists of four studies that investigated speaker idiosyncrasy and intensity variability in the speech signal. Sound intensity refers to the acoustic power per unit area, and is often calculated as the logarithm of the ratio where an arbitrary acoustic power serves as the reference. Such measurements are often referred to as *intensity levels*. In this dissertation, I use "intensity" to mean such logarithmic ratio measures. Related physical concepts such as amplitude, magnitude, pressure, energy and power, as well as psychophysical loudness measures are easily confusable with intensity. In the appendices to this thesis, I explain in more details how these concepts are related to each other, and what happens behind the scene when doing intensity analysis using the Praat software (Boersma and Weenink 1992-2016). In the upcoming paragraphs of this synopsis, I introduce the background of my research, and summarize the purposes and findings of the four empirical studies that cumulatively form this thesis.

Research background

This thesis evolved phonetic from research on speech rhythm. Traditionally, speech rhythm was defined in terms of isochrony (Abercrombie 1967; Lloyd James 1940; Pike 1945): Germanic languages such as English, German, and Dutch were believed to have isochronous feet (i.e., equal between-stress interval durations), hence were referred to as "stress-timed" languages. Romance languages such as French, Italian and Spanish were believed to have isochronous syllables (i.e., equal syllable durations), hence were referred to as "syllable-timed" languages. For "stress-timed" languages like English, it is possible to squeeze a number of unstressed syllables between stressed syllables, without dramatically changing the duration of the foot. However, for "syllable-timed" languages like French, syllables – stressed or unstressed – appear to have similar durations. Such tendency of isochrony based on different linguistic units makes it possible to illustrate the rhythmic characteristics between "stress-" and "syllable-timed" languages using music notations (**Figure S-1**).

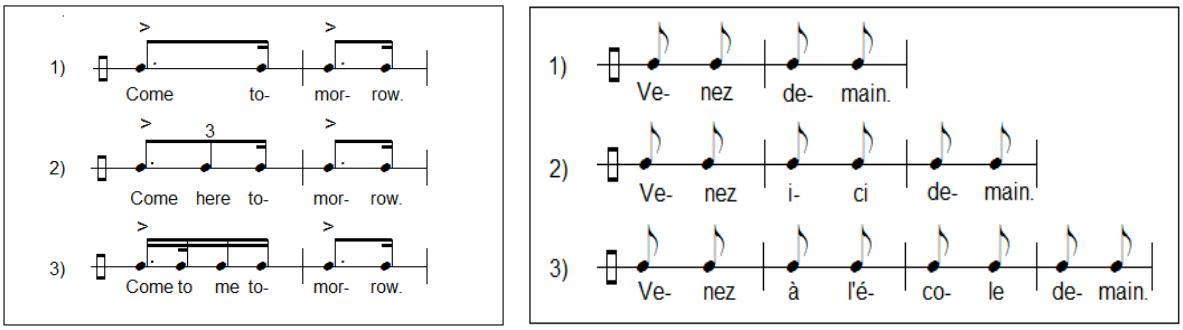


Figure S-1: Music notations showing the tendency of isochronous feet for English (left) and isochronous syllables for French (right).

Nevertheless, such perfect isochrony can easily break down in spontaneous speech. Early instrumental measurements of foot and syllable durations failed to show such isochronous patterns (e.g., Bertrán 1999; Dauer 1983; Pointon 1980; Roach 1982). For this reason, Nespor (1990) even rejected "stress-" and "syllable-timing" as legitimate metalinguistic terms. However, a myriad of studies have shown that languages differing in rhythmicity are perceptually salient among adults, neonates and even animals (Bosch and Sebastián-Gallés 1997; Nazzi et al. 1998; Nazzi et al. 2000; Ramus and Mehler 1999; Ramus et al. 1999; Ramus et al. 2000). What makes languages different in rhythmicity is still remaining to be fully understood.

More recently, researchers (Nolan and Jeon 2014) considered speech rhythm as a metaphor. They drew an analogy between speech rhythm and the landscapes with crop fields as observed from an airplane. People often describe what they see as chessboards, although such landscapes rarely manifest perfect square shapes. This is due to the human cognitive ability of metaphorical extension. In terms of speech, listeners perceive the acoustic events as manifesting some regularly occurring patterns in the absence of strict beats. Similarly in music, composers often mark the fermata symbol (𝄐) over a particular note in the score so that players, singers, or conductors can prolong the duration of a particular note to make the passage more expressive. This disrupts the assumed duration of the note and the bar where this prolonged note belongs to. Listeners, nevertheless, never fail to appreciate the rhythmicities of the work.

Since perfect isochronous feet or syllables were rarely found in speech by acoustic measurements, researchers started off to investigate the durational

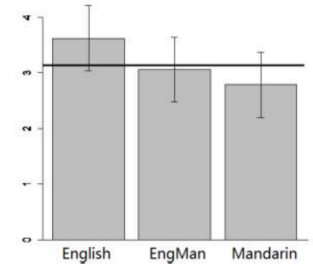
characteristics of other units in the speech signal and proposed a number of metrics to quantify rhythmic differences between languages. Dauer (1983, 1987) noted that languages showing different rhythmicity differ in terms of syllable weight and vowel reduction: "stress-timed" languages usually have complicated syllable structures and a higher degree of vowel reductions, whereas "syllable-timed" languages have simpler syllable structures with a lower degree of vowel reductions. Ramus et al. (1999) quantified this idea by measuring the standard deviation of consonantal interval durations and vocalic interval durations (ΔC and ΔV), as well as the percentage of vocalic interval durations (%V) sentencewise. Grabe and Low (2002) applied the pairwise variability indices (PVI) to measure the durational variability of consecutive vocalic or intervocalic intervals (nPVI_V and rPVI_C). Both measures segregated languages of different rhythmicity. Additionally, Dellwo (2009, 2010) developed a number of normalization methods (e.g., variation coefficient (varco) and natural logarithm) to counteract the effect of speech rate variability. In the meanwhile, other approaches to speech rhythm also emerged, such as the coupled-oscillator model (O'Dell and Nieminen 1999), amplitude modulation phase model (Leong et al. 2014), auditory primal sketch model (Lee and Todd 2004), and the spectral model of the amplitude envelope (Tilsen and Johnson 2008). They did not measure duration directly from the signal, but all searched for temporal regularities in the signal. The introductory section of Study III (§1.1) offers a more extensive review of these approaches.

This thesis innovates speech rhythm research in that it focuses on intensity, another important aspect in the rhythmicitity of the speech signal. Why should intensity play a role in between-language rhythm difference? 1) Differences in phonotactic structures should be the source of intensity variability between languages. This is similar to durational cues to speech rhythm. For example, open vowels are intrinsically louder, and hence carry more intensity than closed vowels (see Lehiste and Peterson 1959 for a survey of intrinsic intensities of American English vowels). Experience also shows that consonants carry different intensities. Klatt (1980) and Coleman and Slater (2001) estimated different amplitude parameters for synthesizing stops, fricatives and affricates and other sonorant consonants using the Klatt synthesizer (Klatt 1980; Klatt and Klatt 1990). This reflects

the fact that in order for the synthesized consonants to sound natural, it is essential to modify and regulate amplitudes in different frequency bands, which in turn affects the overall segmental intensity levels. Languages differ in their segment inventories, and rules of segment combinations. A phonotactically more complex language, such as English and German should have higher levels of intensity variability than Italian or French as calculated from mean syllabic intensity levels across syllables.

2) Reduced or centralized vowels are not only shorter in duration, but also lower in amplitude levels, and therefore should have lower intensity levels in terms of mean intensity or peak intensity. Languages that allow vowel reductions should have higher syllabic intensity variability. Similarly, language-specific stress characteristics should also result in different patterns of intensity variability. For languages (e.g., English) using intensity as a cue to stress or prominence, intensity variability should be higher than the ones that do not have lexical stress or do not rely on intensity information to signal stress (e.g., Mandarin, see Wang 2008). It has been found that second language learners' (L1 = Mandarin, L2 = English) speech exhibited a pattern of intensity variability similar to Mandarin (He 2012), even though the pattern of duration variability is rather similar to English (He 2010). Both He (2010) and He (2012) predated my PhD research; however, they play a significant role in the conceptualization and development of the method used in my PhD research. Therefore, the following two paragraphs summarize the major findings of these two studies.

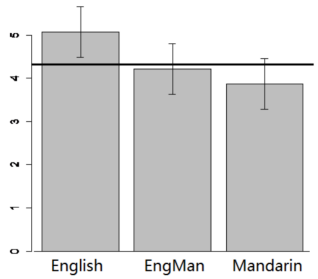
• He (2010): I measured the rhythm of native American English (Abbreviated as L1 English henceforth), native Beijing Mandarin (Abbreviated as L1 Mandarin henceforth), and the second language English of Mandarin speakers (Abbreviated as Eng_{Man}) using the measures of ΔC, ΔV, %V, varcoC, varcoV, rPVI_C and nPVI_V for my master's thesis at the University of Edinburgh (He 2010; partial results were also published in He 2014). The differences between L1 English and L1 Mandarin were significant as expected. However, L1 English and EngMan were not significantly different, although they were impressionistically different in terms of rhythmicity. Similar results were also obtained by Mok and Dellwo (2008).



	Df	Sum Sq	Mean Sq	F
Languages	2	9.092	4.5458	7.7748***
Residuals	72	42.097	.5847	
Total	74	51.189		

*** $p < .0001$

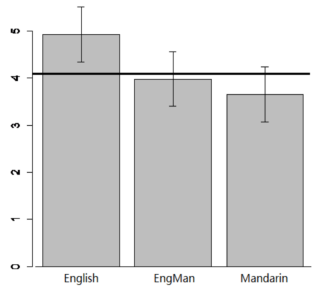
standard deviation



	Df	Sum Sq	Mean Sq	F
Languages	2	19.112	9.5560	7.6716***
Residuals	72	89.686	1.2456	
Total	74	108.798		

*** $p < .0001$

variation coefficient



	Df	Sum Sq	Mean Sq	F
Languages	2	21.763	10.8816	6.1966**
Residuals	72	126.437	1.7561	
Total	74	148.200		

** $p < .001$

PVI

Figure S-2: Bar plots and Fisher statistics showing the differences between three measures of average syllabic intensity variability in L1 English, Eng_{Man} and Mandarin. Error bars indicate ±1 std.err; the horizontal lines indicate pooled means. Plots and statistical results were originally published in He (2012).

• He (2012): Enlightened by Wang (2008), who showed that native speakers of English and Mandarin differed in the use of intensity as stress cues (for English listeners intensity is a salient stress cue, but not for Mandarin listeners), I hypothesized that intensity variability is smaller in Eng_{Man} than in L1 English. The impressionistic rhythmic difference between L1 English and EngMan may be due to different degrees of intensity variability in the speech signal. I therefore measured the standard deviation, variation coefficient and pairwise variability index of the average syllabic intensity using the same database as in He (2010, 2014), and found that both Eng_{Man} and L1 Mandarin were significantly lower than L1 English in terms of intensity variability (see Figure S-2).

During my PhD studies, I was working in the project of "VoiceTime" (http://www.grstiftung.ch/de/portfolio/projekte/alle/y_2013/GRS-027-13.html)

that aims to implement rhythmic information to automatic speaker recognition systems. This project is a sister project to the completed "Forensic Phonetic Speaker Identification Based on Temporal Evidence" (http://www.research-projects.uzh.ch/p15317.htm), which already showed that speaker idiosyncrasy is reflected in terms of durational measures of rhythm in different modalities of speech: read speech (Dellwo et al. 2012, 2015; Leemann et al. 2014), spontaneous speech (Dellwo et al. 2015; Leemann et al. 2014), GSM transmitted speech (Leemann et al. 2014), L2 speech (Dellwo and Schmid 2015), disguised speech (Leemann and Kolly 2015), and speech differed highly in rate (Dellwo et al. 2015). The rationale behind these studies is that humans differ in terms of the anatomical dimensions of the articulators, which results in idiosyncratic temporal characteristics in articulation, namely speaker-specific rhythm (Dellwo et al. 2015; Leemann et al. 2014). I believe that not only can such anatomical idiosyncrasy be measurable in duration-based rhythm measures, but also in intensity-based rhythm measures (see the Introduction of Study III for a more detailed exposition of rationale). Studies I, II and III explored speaker-specific speech rhythm using intensity-based rhythm measures. Study IV moved beyond speech rhythm and explored dynamic characteristics of intensity fluctuations in the signal. Such intensity dynamics are closely related to articulatory movements, especially those having directly influence on the area of mouth opening. The introduction of Study IV gives a detailed exposition of the rationale. Moreover, how speaker idiosyncrasy is manifested in intensity dynamics is also explored in Study IV. The purposes, results, and how these four studies are connected to each other are summarized below:

Study I

Study I provides the first evidence that between-speaker intensity variability is significant using the TEVOID database designed to explore between-speaker rhythmic variability. More measures of intensity variability (referred to as intensity

measures hereafter) were tested in this study (as well as in Studies II and III) than in He (2012). In addition to average syllabic intensity, the peak intensity for each syllable was measured as well (see Figure S-3). The standard deviation, variation coefficient (i.e., normalized standard deviation, also abbreviated as varco), raw PVI and normalised PVI were calculated for both average syllabic intensity (stdevM, varcoM, rPVIm and nPVIm, collectively referred to as mean measures) and syllable peak intensity (stdevP, varcoP, rPVIp and nPVIp, collectively referred to as peak measures). Formulas for these measures are listed in §2.2 of study I and Table 1 of study III (page 54). Rationale for these measures were explicated in the introduction of study III (page 47).

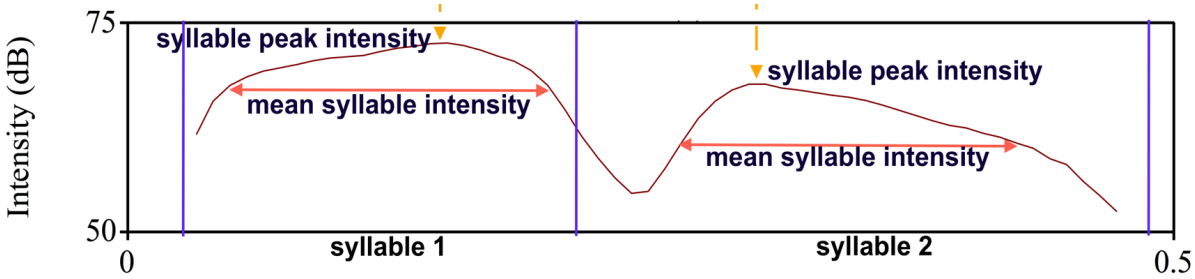


Figure S-3: An illustration of the analysing units (mean syllabic intensity & syllable peak intensity) for the calculation of intensity measures.

For the statistical procedures, pairwise correlations of these intensity measures were evaluated, and univariate ANOVA models (dependent variables: intensity measures; independent variable: speaker) were tested. The following results were obtained:

- Raw measures and normalized measures were highly correlated;
- Mean measures and peak measures were poorly correlated;
- Between-speaker variability was significant for all intensity measures.

Study II

Since between-speaker variability was significant for all intensity measures, it would be interesting to test how well automatic speaker recognition algorithms perform based on these intensity measures. Also, it would be interesting to compare

intensity measures with duration measures to see which set of measures are better at recognizing speakers. Duration measures tested in Leemann et al. (2014) were applied. The artificial neural networks (ANN) were used for the recognition experiments with the TEVOID database. Prior to training the ANNs, the *k*-nearest neighbours (*k*NN) algorithm was tested on intensity measures first, but the performance was not satisfactory: no more than 12% hit-rate was obtained with *k*NNs (He et al. 2014). With the ANN, the recognition rates in general increased. The major findings of study II were summarised as follows:

- The mean recognition rates obtained from the test set were 14.2% (duration only), 30.3% (intensity only), and 36.9% (duration cum intensity);
- The recognition rates using intensity measures and combined measures were significantly higher than those using duration measures alone;
- The intensity measures and combined measures were not significantly different in recognising speakers.

The implications for such findings are, 1) intensity measures performed better in recognizing speakers than duration measures, and 2) the choice of recognition algorithms is important. These points were also tested in more details in Study III.

Study III

Study III elaborates the ideas of the previous two studies with more detailed analyses. In addition to the TEVOID database, the BonnTempo database (high within-speaker variability) was also used. This study was conducted to explore 1) which are the intensity measures that best account for between-speaker variability in both databases, and 2) how well intensity measures and duration measures perform to recognize speakers, which includes the following sub-points:

- Which domain of measures, intensity, duration, or combined, provide higher speaker recognition results?

- Are there any effect of recognition algorithms (artificial neural network vs. multinomial logistic regression) on recognition results?
- Are there any effect of the z-score transformation by sentence on speaker recognition performance?

The following results were obtained:

- The measures of standard deviations and variation coefficients (stdevM/P, varcoM/P) explained more between-speaker variability than the PVI measures. In addition, the peak measures (stdevP, varcoP, rPVIp and nPVIp) conjointly explained more between-speaker variability than the mean measures (stdevM, varcoM, rPVIm and nPVIm);
- Intensity measures and combined measures outperformed duration measures alone in speaker recognition;
- A significant effect of the recognition algorithm was found: multinomial logistic regression outperformed artificial neural network in recognizing speakers using the rhythm measures;
- A significant effect of the z-score transformation on speaker recognition performance was found: z-score transformed measures outperformed non-normalized measures.

Study IV

Study IV moves beyond intensity measures and speaker-specific speech rhythm to investigate speaker idiosyncratic intensity dynamics. Intensity dynamics is defined as the speed of an intensity increase from the amplitude envelope trough to the peak (positive dynamics), or the speed of an intensity decrease from the peak to the trough (negative dynamics). Both dynamics are geometrically illustrated in Figure S-4.

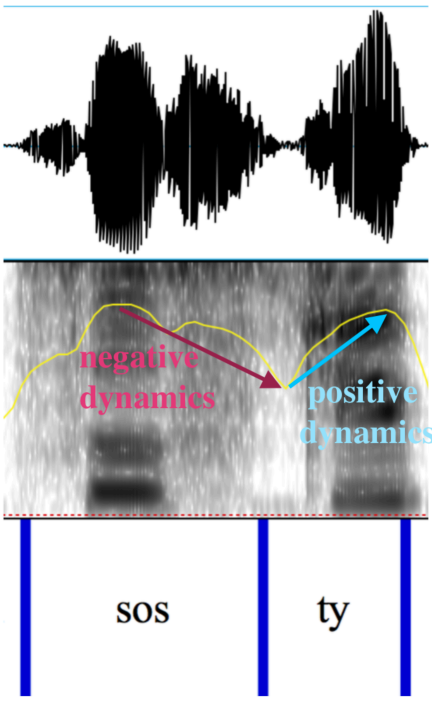


Figure S-4: An geometric illustration of positive intensity dynamics and negative intensity dynamics.

The motivation to investigate speaker idiosyncratic intensity dynamics is twofold: 1) since intensity and duration measures of speech rhythm show significant between-speaker variability (Studies I, II, and III; Dellwo et al. 2015; Leemann et al. 2014), it is also interesting to integrate intensity and duration to study speaker idiosyncrasy in terms of intensity change as a function of time. 2) More importantly, it has been discovered that the intensity contour shape is closely related to the articulatory movements repsonsible for the changes of mouth opening (Chandrasekaran et al. 2009). Speaker-specific articulatory behaviors should therefore be reflected in the dynamics of intensity contours. Morever, by examining the coordination between articulators, it was discovered that speakers are likely to have different motor planning for opening and closing gestures (Birkholz et al. 2011). Therefore, the study tested speaker-specific variability in both positive and negative dynamics. It was discovered that between-speaker variabiliy was largely explained by measure of negative dynamics. This suggests that the closing gestures of articulatory movements might be more speaker-specific.

Summary

This thesis investigates speaker idiosyncrasy in intensity variability of the speech signal. A set of intensity measures were developed as an alternative approach to quantifying speech rhythm. The focus of this work is on between-speaker rhythmic differences. The first three studies have shown that between-speaker variability was significant as measured by intensity-based rhythm measures. Moreover, they outperformed duration-based rhythm measures in automatic speaker recognition experiments. Additionally, a set of measures of intensity dynamics were created in Study IV. It was discovered that measures of negative dynamics explained more between-speaker variability than positive dynamics, suggesting that speaker individual differences might be largely encoded in the closing gestures of articulatory movements.

References

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh, UK: Edinburgh University Press.

Bertrán, A. P. (1999). Prosodic typology: On the dichotomy between stress-timed and syllable-timed languages, *Language Design* 2: 103-131.

Birkholz, P., Kröger, B. J. and Neuschaefer-Rube, C. (2011). Model-based reproduction of articulatory trajectories for consonant-vowel sequences. *IEEE Transactions on Audio, Speech, and Language Processing* 19: 1422-1433.

Boersma, P. and Weenink, D. (1992-2014). Praat: doing phonetics by computer, Retrieved from http://www.praat.org/

Bosch, L. and Sebastián-Gallés, N. (1997). The role of prosody in infants' native language discrimination abilities: The case of two phonologically close languages, in *EUROSPEECH 1997*, Rhodes, Greece, pp. 231-234.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A. and Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology* 5: e1000436.

Coleman, J., and Slater, A. (2001). Estimation of parameters for the Klatt synthesizer for a speech database, in R. I. Damper (ed.), *Data-Driven Techniques in Speech Synthesis*, Dordrecht: Springer Science+Business Media, pp. 215-238.

Dauer, R. (1983). Stress-timing and syllable-timing reanalyzed, *Journal of Phonetics* 11: 51-62.

Dauer, R. (1987). Phonetic and phonological components of language rhythm, in *International Congress of Phonetic Science (ICPhS) XI*, Tallinn, Estonia, pp. 447-450.

Dellwo, V. (2009). Choosing the right rate normalization method for measurements of speech rhythm, in S. Schmid, M. Schwarzenbach and D. Studer (eds.), *La Dimensione Temporale del Parlato: Atti del 5° Convegno Nazionale AISV 2009*, Torriana, Italy: EDK Editore, pp. 13-32.

Dellwo, V. (2010). *Influences of Speech Rate on the Acoustic Correlates of Speech Rhythm: An Experimental Phonetic Study Based on Acoustic and Perceptual Evidence*, Doctoral dissertation, Bonn University, Germany.

Dellwo, V., Leemann, A. and Kolly, M.-J. (2012). Speaker idiosyncratic rhythmic features in the speech signal, in *Interspeech 2012*, Portland (OR), USA, pp. 1582-1585.

Dellwo, V., Leemann, A., and Kolly, M.-J. (2015). Rhythmic variability between speakers: articulatory, prosodic, and linguistic factors, *Journal of the Acoustical Society of America* 137: 1513-1528.

Dellwo, V. and Schmid, S. (2015). Speaker-individual rhythmic characteristics in read speech of German-Italian bilinguals, in A. Leemann, M.-J. Kolly, S. Schmid and V. Dellwo (eds.), *Trends in Phonetics and Phonology: Studies from German-Speaking Europe*. Bern: Peter Lang, pp. 349-362.

Grabe, E. and Low E. L. (2002). Durational variability in speech and rhythm class hypothesis, in C. Gussenhoven and N. Warner (eds.), *Laboratory Phonology 7*. Berlin, Germany: Mouton de Gruyter, pp. 514-546.

He, L. (2010). *Interlanguage Rhythm: A Durational Metrics Study among Native Speakers of Mandarin and Cantonese Learning English*, MSc dissertation in developmental linguistics, University of Edinburgh.

He, L. (2012). Syllabic intensity variations as quantification of speech rhythm: Evidence from both L1 and L2, in *Speech Prosody 2012*, Shanghai, China, pp. 466-469.

He, L. (2014). The inadequacy of rhythm metrics to quantify L2 suprasegmental characteristics, in *Speech Prosody 2014*, Dublin, Ireland, pp. 1095-1099.

He, L., Glavitsch, U. and Dellwo, V. (2014). Automatic speaker identification using syllable intensity variability: an initial attempt using the *k*NN classifier. Abstract presented at Phonetik & Phonologie 10, Konstanz, Germany. http://ling.uni-konstanz.de/pages/conferences/pp10/abstracts/He_pp10.pdf

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer, *Journal of the Acoustical Society of America* 67: 971-995.

Klatt, D. H. and Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers, *Journal of the Acoustical Society of America* 82: 820-857.

Lee, C. S. and Todd, N. P. M. (2004). Towards an auditory account of speech rhythm: application of a model of the auditory "primal sketch" to two multi-language corpora, *Cognition* 93: 225-254.

Leemann, A., Kolly, M.-J. and Dellwo, V. (2014). Speech-individuality in suprasegmental temporal features: implications for forensic voice comparison, *Forensic Science International* 238: 59-67.

Leemann, A. and Kolly, M.-J. (2015). Speaker-invariant suprasegmental temporal features in normal and disguised speech, *Speech Communication* 75: 97-122.

Lehiste, I. and Peterson, G. E. (1959), Vowel Amplitude and Phonemic Stress in American English, *Journal of the Acoustical Society of America* 31: 428-435.

Leong, V., Stone, M. A., Turner, R. E. and Goswami, U. (2014). A role for amplitude modulation phase relationships in speech rhythm perception, *Journal of the Acoustical Society of America* 136: 366-381.

Lloyd James, A. (1940). *Speech Signals in Telephony*. London: Sir Isaac Pitman & Sons.

Mok, P. and Dellwo, V. (2008). Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English, in *Speech Prosody 2008*, Campinas, Brazil, pp. 423-426.

Nazzi, T., Bertoncini, J. and Mehler, J. (1998). Language discrimination by newborns: Towards an understanding of the role of rhythm, *Journal of Experimental Psychology: Human Perception and Performance* 24: 756-766.

Nazzi, T., Jusczyk, P. W. and Johnson, E. K. (2000). Language discrimination by English-learning 5-month-olds: Effect of rhythm and familiarity, *Journal of Memory and Language* 43: 1-19.

Nespor, I. (1990). On the rhythm parameter in phonology, in I. Roca (ed.), *Logical Issues in Language Acquisition*. Dordrecht: Foris, pp. 157-195.

Nolan, F. and Jeon, H.-S. (2014). Speech rhythm: a metaphor? *Philosophical Transactions of the Royal Society B* 369: 20130396.

O'Dell, M. L. and Nieminen, T. (1999). Coupled oscillator model of speech rhythm, in *International Congress of Phonetic Sciences (ICPhS) XIV*, San Francisco, pp 1075-1078.

Pike, K. (1945). *The Intonation of American English*. Ann Arbor: University of Michigan Press.

Pointon, G. E. (1980). Is Spanish really syllable-timed? *Journal of Phonetics* 8: 293-304.

Ramus, F. and Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis, *Journal of the Acoustical Society of America* 105: 512-521.

Ramus, F., Nespor, M. and Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal, *Cognition* 73: 265-292.

Ramus, F., Hauser, M. D., Miller, C., Morris, D. and Mehler, J. (2000). Language discrimination by human newborns and by cotton-top Tamarin monkeys, *Science* 288: 349-351.

Roach, P. (1982). On the distinction between "stress-timed" and "syllable-timed" languages, in D. Crystal (ed.), *Linguistic Controversies*. London: Edwards Arnold, pp. 73-79.

Tilsen, S. and Johnson, K. (2008). Low-frequency Fourier analysis of speech rhythm, *Journal of the Acoustical Society of America* 124: EL34-EL39.

Wang, Q. (2008). L2 stress perception: The reliance on different acoustic cues, in *Speech Prosody 2008*, Campinas, Brazil, pp. 635-638.

■

Study I

Speaker idiosyncratic variability of intensity across syllables

This work has been published in September 2014 in the *Proceedings of INTERSPEECH 2014*, Singapore. pp. 233-237.
Official URL:
http://www.isca-speech.org/archive/archive_papers/interspeech_2014/i14_0233.pdf

Speaker Idiosyncratic Variability of Intensity across Syllables

Lei He, Volker Dellwo

Phonetics Laboratory, University of Zurich, Switzerland

{lei.he, volker.dellwo}@uzh.ch

Abstract

This study explored speaker idiosyncrasy by measuring the syllabic intensity variability in the speech signal. Sixteen speakers of the TEVOID corpus, each producing 256 read sentences, were analyzed. Characteristics of intensity variability (average or peak) between syllables were measured either holistically (standard deviation of intensity changes between syllables) or locally (pairwise variability indices of intensity changes between syllables). The results indicated significant effects of the speakers in all the metrics, suggesting a potential application of the methods for speaker recognition, and in particular for forensic speaker comparison.

Index Terms: intensity variability, speaker idiosyncrasy

1. Introduction

Speech production is a complicated process that involves much neuromuscular programming to control the movements of articulators [1]. The motor control in speech, similar to other modes of human movements like human gain [2, 3], is highly individual, and it seems conceivable that such individual characteristics are reflected in the physical properties of the speech signal. Enlightened by the idiosyncratic temporal characteristics in human gait, the research team in our laboratory adopts a time-domain approach to voice identification. The widely used speech rhythm metrics [4, 5, 6, 7] were employed to find speaker individualities in the speech signal. [8] and [9] discovered that the percentages over which speech is vocalic (%V) and the percentage over which speech is voiced (%VO) showed fair success in detecting speaker idiosyncrasy with spontaneous speech. %VO also turned out to show speaker specific characteristics independent of the language in bilingual speakers [10]. Moreover, newly developed metrics (ΔPeak) also succeeded in finding speaker individualities [8, 9]. ΔPeak is calculated by taking the standard deviations of the intervals between syllabic amplitude peaks. Such measures are motivated by the idea that the combined movements of the articulators result in a temporal organization of amplitude envelope characteristics like syllabic peak points. This idea also motivates the present study in which we studied amplitude peak and syllabic intensity variability between speakers. Similar to temporal measures we previously found that such measures show language specific effects between English and Mandarin or L2 English by Mandarin natives [11]. In the present study we tested to what degree such measures reveal within-language variability as a function of speakers, if speaker specific controls of the articulators are responsible for the individual timing organization of speech.

2. Methods

2.1. The TEVOID corpus

The TEVOID (Temporal Voice Idiosyncrasy) corpus [8, 9] was constructed in the Phonetics Laboratory of the University of Zurich to study temporal variability in the speech signal. The speakers were all native speakers of Zurich German. This German variety shows little if any socio-economic variability, which could be a potential artifact in between-speaker variability of temporal characteristics [8]. Recordings of both read and spontaneous speech of 16 speakers are in the current corpus. All the recordings were digitized in a sound attenuated booth with the sampling rate of 44.1 kHz and a quantization depth of 16 bit. The read speech (256 sentences * 16 speakers = 4'096 sentences) was analyzed for the present study.

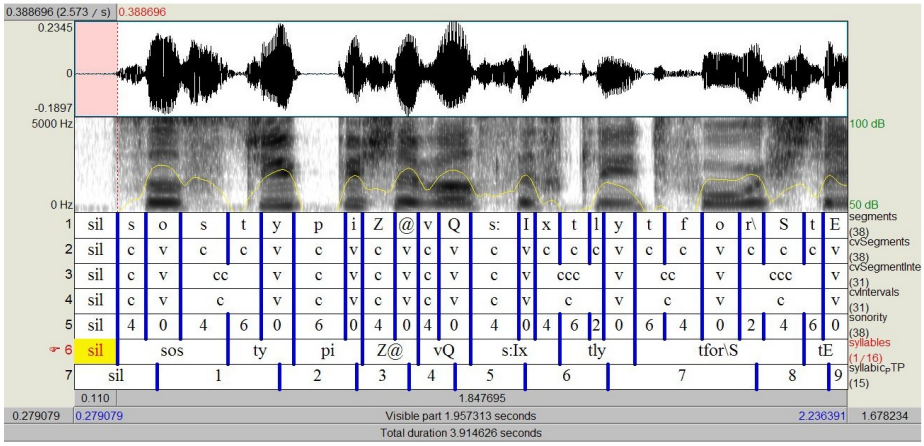


Figure 1. *A screenshot of the TEVOID annotations. Tier 6 and tier 7 mark the syllables and the corresponding syllable peaks in the utterance.*

Segments of the sound files were annotated manually, on the basis of which tiers of different interval details were created. However, for the present study, only the tier containing syllable on- and offset information (Tier 6, Figure 1) and the tier containing the times of syllabic amplitude peak points (Tier 7, Figure 1) were relevant for the measurements described below (§2.2). For more about the TEVOID corpus, please refer to [8, 9].

2.2. Techniques of measurement

Two sets (mean metrics and peak metrics) of intensity variability measures were devised based on the durational metrics of speech rhythm (ΔC/V, varcoC, rPVI, nPVI) [4, 5, 6]. The basic calculation unit of the mean metrics is the average intensity across each syllable in the utterance, while the calculation unit of the peak metrics is the intensity of each automatically detected syllable peak. For both sets, global intensity variations are quantified by taking the standard deviations of both mean and peak intensity of the syllables in a sentence utterance, hence the metrics stdevM and stdevP. Local intensity variations are quantified by taking the cumulative intensity differences between adjacent syllables, either in the mean tier or peak tier. Formulas (1) and (2) express the idea more explicitly:

$$\text{rPVIm} = \sum_{j=1}^{n-2} | L_{\text{M}j} - L_{\text{M}j+1} | / (n-2) \quad (1)$$

$$\text{rPVIp} = \sum_{j=1}^{n-2} | L_{\text{P}j} - L_{\text{P}j+1} | / (n-2) \quad (2)$$

where rPVI means the raw pairwise variability index; $L_{\text{M}j}$ and $L_{\text{P}j}$ refer to the mean intensity level and the peak intensity level of the jth syllable in the utterance; and n refers to the total number of intervals in the utterance.

However, some speaker may be intrinsically "louder" than others, and the distance between the mouth and the microphone may not be precisely controlled either. Hence the global metrics are normalized by taking the ratios of the original scores and the average intensity levels in the mean and peak tiers: varcoM = 100 * stdevM / $\bar{L}_{\text{M}}$, varcoP = 100 * stdevP / $\bar{L}_{\text{P}}$, where varco is short for variation coefficient; $\bar{L}_{\text{M}}$ and $\bar{L}_{\text{P}}$ refer to the average syllabic intensity levels in the mean and peak tiers. The local measures are normalized by dividing the absolute difference of each neighboring pair by their own average value prior to the final summation:

$$\text{nPVIm} = \sum_{j=1}^{n-2} \left| \frac{L_{\text{M}j} - L_{\text{M}j+1}}{[L_{\text{M}j} + L_{\text{M}j+1}]/2} \right| \times \frac{100}{n-2} \quad (3)$$

$$\text{nPVIp} = \sum_{j=1}^{n-2} \left| \frac{L_{\text{P}j} - L_{\text{P}j+1}}{[L_{\text{P}j} + L_{\text{P}j+1}]/2} \right| \times \frac{100}{n-2} \quad (4)$$

where nPVI is the normalized pairwise variability index, and the denotations of the other symbols are the same as those in formulas (1) and (2). The scalar 100 in both varcoM, varcoP as well as in (3) and (4) makes the integer parts of the scores greater than zero.

The calculations were automated in Praat [12] using a script (available from the first author). The parameters for extracting and querying the intensity objects were set default (minimum pitch = 100 Hz, time steps = 0.0, subtract mean = True, averaging method = dB, interpolation method = Cubic). The initial and final syllables were excluded from analysis because the duration was sometimes too short for the intensity values to be measured reliably.

3. Data analysis and results

3.1. Data normality and transformations

Data distributions were assessed by constructing Q-Q plots for all the metrics. The data deviate from the normal Q-Q lines as the top left panel of Figure 2 (only rPVIp is displayed due to limited space, but the patterns are similar across metrics) shows, not meeting the distribution assumption of parametric statistics. Therefore, natural logarithmic transformations ($X_{\text{Trans}} = \ln X$), square root transformations ($X_{\text{Trans}} = X^{1/2}$) and arcsine transformations ($X_{\text{Trans}} = (2/\pi) * \sin^{-1}(X/100)^{1/2}$) [13] were performed to see which methods optimally transform the data into normally distributed ones. As the Q-Q plots in Figure 2 indicate, the natural log transformations had the least success, whereas the square root and arcsine transformations showed similar success in transforming the data into normally distributed sets. Given similar effects of both square root and arcsine transformations, we will only analyze the square root transformed data in the coming sections, because the calculations are more straightforward compared with the arcsine transformations.

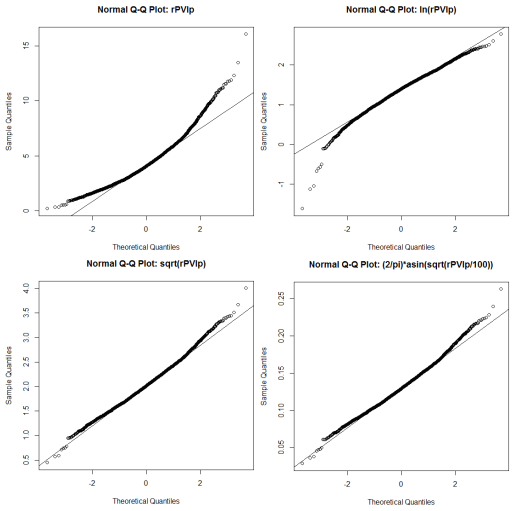


Figure 2. *Normal Q-Q plots of the original rPVIp (top left), natural log transformed rPVIp (top right), square root transformed rPVIp (bottom left), and arcsine transformed rPVIp (bottom right).*

3.2. Correlations between the metrics scores

First of all, the correlations between the raw and normalized metrics (stdevM/P vs. varcoM/P; rPVIm/p vs. nPVIm/p) were evaluated to see how well one metric can predict another. Figure 3 demonstrates the scatter plot matrices of both mean and peak metrics. It is evident that the raw scores are highly correlated with their normalized counterparts. The Pearson's correlation coefficients confirm that very high correlations exist: $r = 0.974$ for stdevM and varcoM, $r = 0.981$ for rPVIm

and nPVIm, $r = 0.991$ for stdevP and varcoP, and $r = 0.993$ for rPVIp and nPVIp (all p values < 0.0001).

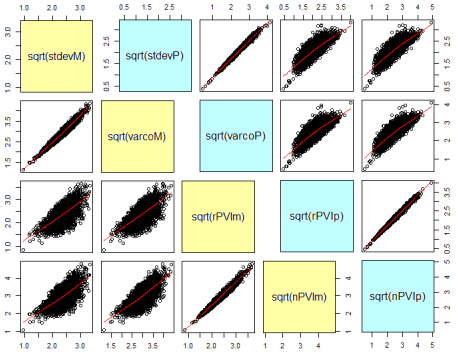


Figure 3. *The scatter plot matrices of both mean metrics (lower panel) and peak metrics (upper panel). The red LOWESS lines indicate that the relationships are linear to a large extent, and the raw scores and their normalized counterparts are highly correlated.*

In addition, correlations between the holistic and local metrics within the mean and peak measures are also significantly high: $r = 0.666$ for stdevM and rPVIm, $r = 0.662$ for varcoM and rPVIm, $r = 0.670$ for stdevM and nPVIm, and $r = 0.707$ for varcoM and nPVIm (all p values < 0.0001); $r = 0.794$ for stdevP and rPVIp, $r = 0.784$ for varcoP and rPVIp, $r = 0.804$ for stdevP and nPVIp, and $r = 0.809$ for varcoP and nPVIp (all p values < 0.0001).

Finally, we also examined the correlations between the mean metrics and the peak metrics. As the scatter plots (Figure 4) show, the correlation of both holistic metrics and local metrics between the mean and peak measures are rather poor (the highest correlation coefficient being merely 0.322 for nPVIm and nPVIp, $p < 0.0001$). This means that mean and peak measures cannot be reliably predicted from each other, suggesting that both measures contain different information about the amplitude contour.

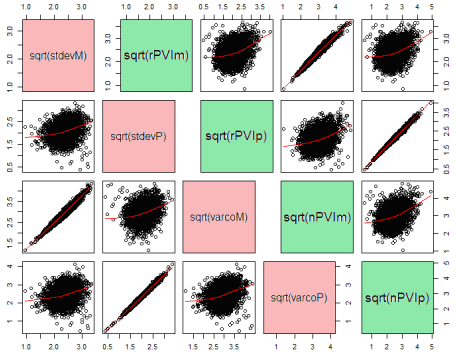


Figure 4. *The scatter plot matrices of both holistic measures (lower panel) and local measures (upper panel) between the mean and peak measures. The red LOWESS lines also indicate that the poorly correlated scores deviate from linearity to some extent.*

3.3. Significant effects of speakers

Figures 5 to 8 contain box plots for the measures of varcoM, nPVIm, varcoP, and nPVIp as a function of the 16 speakers. It is visible that there are strong differences between the speakers. For some of the measures the between speaker variability seems to be rather similar (nPVIp and varcoP), but not for the other measures. Univariate ANOVAs with speakers as the independent variable were performed on the square root transformed data in R [14]. For each metric, the Bartlett test of variances homogeneity was run, so as to adjust the "var.equal" argument in the ANOVA commands. If the equality of variances is violated, an approximate method of Welch [15] is applied in the computation. As Table 1 shows, significant effects of the speakers were found on all the metrics, both mean measures and peak measures.

Table 1. *Statistical outputs of the square root transformed data (var.equal=FALSE for all metrics as indicated by K^2).*

	Bartlett's K^2 (*df*)	*F*	*df* (num, denom)
stdevM	82.21⋆ (15)	47.43⋆	15, 1540.69
varcoM	109.30⋆ (15)	66.53⋆	15, 1540.17
rPVIm	85.46⋆ (15)	32.08⋆	15, 1540.74
nPVIm	108.93⋆ (15)	43.37⋆	15, 1540.65
stdevP	54.87⋆ (15)	88.25⋆	15, 1540.74
varcoP	60.14⋆ (15)	98.52⋆	15, 1540.72
rPVIp	64.15⋆ (15)	90.23⋆	15, 1540.76
nPVIp	64.94⋆ (15)	95.30⋆	15, 1540.76

⋆$p < 0.0001$

The speaker individualities are also visualized by box plots (Figures 5 – 8). As Figure 3 shows, the raw metrics scores and the normalized ones are highly correlated, thus we only plot the normalized metrics (varcoM, nPVIm, varcoP and nPVIp), because their raw counterparts should have similar patterns.

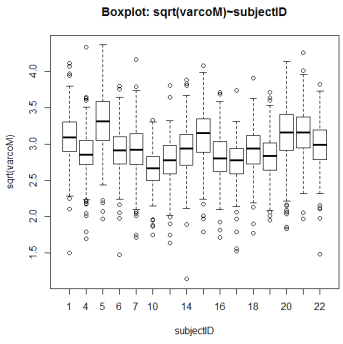


Figure 5. *Box plot of square root transformed varcoM.*

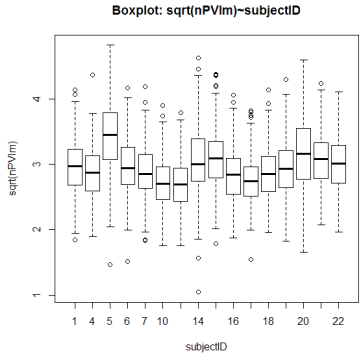


Figure 6. *Box plot of square root transformed nPVIm.*

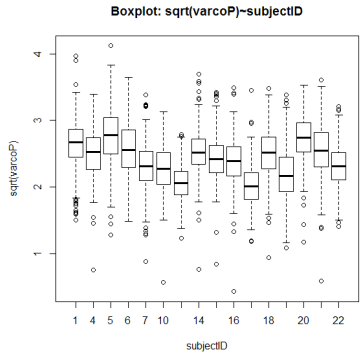


Figure 7. *Box plot of square root transformed varcoP.*

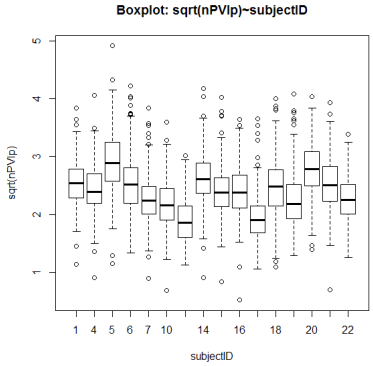


Figure 8. *Box plot of square root transformed nPVIp.*

4. Discussion

This study investigated speaker individualities in the speech signal through variations of intensity levels. Both holistic and local measures of intensity fluctuations (either average syllabic intensities, or syllable peak intensities) were employed. The results indicated that a significant speaker effect exists in all the calculation methods, suggesting a potential application of the methods in speaker recognition.

A closer inspection of the box plots shows that some speakers may not be differentiated from each other on one metric, but this is sometimes not the case on other metrics. For example, speakers 6 and 7 have similar scores on varcoM (Figure 5), but have different ones on varcoP or nPVIp. In a similar vein, speakers 15 and 16 may not be distinguished on varcoP and nPVIp, but they are very likely to be differentiated on varcoM and nPVIm. As Figure 4 suggests, the mean metrics and peak metrics are independent of each, and a combination of the two sets of measures should increase the probability of speaker recognition. Given the fair speaker discriminative strength of durational metrics (%V, %VO, ΔPeak, and varcoPeak) [8], we envisage that speakers can be differentiated even better by a combination of a variety of measures, including the ones presented here.

Moreover, idiosyncratic intensity variability is potentially important for forensic phonetic applications. [16] applied a 1-bit requantization to the speech signal, by setting all positive amplitude values to 1 and all negative ones to 0, thus getting rid of the information of intensity variability contained in the undulating amplitude envelope. However, the 1-bit requantized speech is highly intelligible, suggesting that intensity variability is something in the signal that the speakers may not be aware of. In such situations speakers also have less control over these variables. Therefore, intentional disguise of intensity variability might be more difficult since there is a lack of possible auditory feedback.

For our further research, we would like to see if such idiosyncratic characteristics in intensity levels could also be found in spontaneous speech. In most forensic speaker comparisons, the speech signal is deteriorated to a greater or lesser degree. [17] listed a number of sources of speech degradation, such as reduction of frequency bandwidth, presence of noise, reduction of energy level, spectral distortion, inadequacy of transmission links, and inadequate pickup transducers. Standard audio signal processing techniques like "compressor-limiter" normalization of amplitude levels are non-linear and might introduce a significant amount of noise to between-speakers intensity or amplitude variability. We would like to see if the metrics could, or to what extend, survive these adversities, and how they could be optimized to be useful in actual forensic case works.

5. Conclusion

This study investigated speaker idiosyncrasy via syllabic intensity fluctuations. The results showed that significant effects of the speakers existed in all the intensity metrics, and therefore, are potentially useful in speaker recognition tasks, especially in forensic settings. Future speaker recognition experiments will show whether this hypothesis holds.

6. Acknowledgements

This study is supported by the Gebert-Rüf Stiftung (Grant No. GRS-027/13) and the Swiss National Science Foundation (Grant No. 100015_135287). Adrian Leemann and Marie-José Kolly played a significant role in building the TEVOID corpus. Stephan Schmid made a significant contribution to the conceptualization.

7. References

[1] Dellwo, V., Huckvale, M., and Ashby, M., "How is individuality expressed in voice? an introduction to speech production and description for speaker classification", in C. Müller [Ed], Speaker Classification I, 1-20, Springer Verlag, 2007.

[2] Loula, F., Frasad, S., Kent, H., and Shiffrar, M., "Recognizing people from their movement", J. Exp. Psychol. Hum. Percept. Perform., 31: 210-220, 2005.

[3] Matovski, D., Nixon, M., Mahmoodi, S., and Carter, J., "The effect of time on the performance of gait biometrics", IEEE 4th Conference on Biometrics, Washington, DC, USA, 2010.

[4] Ramus, F., Nespor, M. and Mehler, J., "Correlates of linguistic rhythm in the speech signal", Cognition, 73: 265-292, 1999.

[5] Grabe, E. and Low, E. L., "Durational variability in speech and rhythm class hypothesis", in N. Warner and C. Gussenhoven [Eds], Papers in Laboratory Phonology 7, 515-543, Mouton de Gruyter, 2002.

[6] Dellwo, V., "Rhythm and speech rate: A variation coefficient for deltaC", in P. Karnowski and I. Szigeti [Eds], Language and Language Processing, 231-241, Peter Lang, 2006.

[7] White, L., and Mattys, L. S., "Calibrating rhythm: first language and second language studies", J. Phonet., 35: 501-522, 2007.

[8] Dellwo, V., Leemann, A., and Kolly, M-J., "Speaker idiosyncratic rhythmic features in the speech signal", in Interspeech, Portland, USA, 2012.

[9] Leemann, A., Kolly, M-J., and Dellwo, V., "Speech-individuality in suprasegmental temporal features: implications for forensic voice comparion", Forensic Sci. Int., 238: 59-67, 2014.

[10] Dellwo, V., Schmid, S., Leemann, A., Kolly, M-J., and Müller, M., "Speaker identification based on speech rhythm: the case of bilinguals", Abstract presented at PoRT2012, Glasgow, Scotland, 2012.

[11] He, L., "Syllabic intensity variations as quantification of speech rhythm: evidence from both L1 and L2", in Speech Prosody 6, Shanghai, China, 2012.

[12] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer", version5365, http://www.praat.org/, 2014.

[13] Johnson, K., Quantitative methods in linguistics, Wiley-Blackwell, 2008.

[14] R Core Team, A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/, 2014.

[15] Welch, B. L., "On the comparison of several mean values: an alternative approach", Biometrika, 38: 330-336, 1951.

[16] Kolly, M-J., and Dellwo, V., "Cues to linguistic origin: the contribution of speech temporal information to foreign accent recognition", J. Phonet., 42: 12-23, 2014.

[17] Hollien, H., "About forensic phonetics", Linguistica, 52: 27-53, 2012.

Study II

Comparisons of speaker recognition strengths using suprasegmental duration and intensity variability: an artificial neural networks approach

This work has been published in August 2015 in the *Proceedings of the International Congress of Phonetic Sciences (ICPhS) XVIII*, Glasgow, UK, pp. 0395_1-5.
Official URL:
https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0395.pdf

COMPARISONS OF SPEAKER RECOGNITION STRENGTHS USING SUPRASEGMENTAL DURATION AND INTENSITY VARIABILITY: AN ARTIFICIAL NEURAL NETWORKS APPROACH

Lei He[1], Ulrike Glavitsch[2], Volker Dellwo[1]

[1] Phonetics Laboratory, Department of Comparative Linguistics, University of Zurich, Switzerland
[2] EMPA, Swiss Federal Laboratories for Materials Science and Technology, Dübendorf, Switzerland
lei.he@uzh.ch; ulrike.glavitsch@empa.ch; volker.dellwo@uzh.ch

ABSTRACT

This study compares the speaker recognition strengths based on suprasegmental duration and intensity variability in the speech signal using artificial neural networks. Such algorithm can well capture the nonlinear effects in the data, and is more robust against noise in the data. Three rounds of classification tasks were performed with 1) duration metrics, 2) intensity metrics, and 3) the combination of duration and intensity metrics as the independent variables. The results indicated that both intensity and combined metrics significantly outperformed the duration metrics. Moreover, the combination of intensity and duration metrics showed higher probability of improved speaker classifications than intensity metrics over duration metrics.

Keywords: duration variability, intensity variability, speaker recognition, artificial neural networks

1. INTRODUCTION

Speech production is a complicated process underpinned by sophisticated neuromuscular programming for the motor control of speech organs [6]. The movements of speech organs, like the movements of other parts of the body (see [20, 28] for the example of human gait), are highly idiosyncratic, and such idiosyncrasy should find its acoustic correlates in the speech signal, particularly in the time domain.

In a previous research project of our laboratory [7, 17], the researchers applied the widely used rhythm (duration) metrics (such as ΔC, ΔV, %V, varcoC, varcoV, rPVI-C, and nPVI-V), which were originally developed by [5, 9, 19, 22, 27] to segregate traditionally categorised "stress-" and "syllable-timed" languages [1, 4, 18, 21], and have found significant speaker individualistic temporal characteristics [7, 17].

Along the same line of reasoning, we also hypothesised that individualistic movements as well as anatomical peculiarities of speech organs should result in idiosyncratic energy distribution in the speech signal, and quantifying intensity variability in the signal should capture such idiosyncrasy. Enlightened by the duration metrics, we [2, 10] have developed intensity metrics (please see the appendix) to calculate the syllabic intensity variability (either the mean RMS or peak RMS of each syllable), and the results showed that significant effects of the speaker existed for all the intensity metrics [2, 10].

Our long-term research goal is to explore how successful automatically extracted temporal as well as intensity features will contribute to speaker recognitions, so that they can be implemented in real speaker recognition systems. The present study is an intermediate step towards this goal: we used the human labelled TEVOID corpus (see [7, 17] and §2.1 for more information) and calculated the duration and intensity metrics which were fed into the well-established classification algorithm of artificial neural networks (abbreviated as ANN hereafter), and found that a combination of both duration and intensity metrics gave the best performance of offline speaker recognitions.

The reasons for choosing ANNs were threefold: 1) nonlinear effects in the data, which cannot be controlled for *a priori*, can be modelled by the algorithm [12]; 2) being an eager learner, the ANN generalises the training data before receiving queries from the test data [25], so that the classification is less susceptible to noise; and 3) as a commonly accepted classification algorithm, it can be used as a reference of success for developing new algorithms, which is also in our research pipeline. Primers to the ANN are available as [8, 14, 16], and phonetic research using ANNs include [15, 23, 24, 26], where the latter two focus on speaker recognition.

2. METHOD

2.1. The Corpus

The TEVOID (*Te*mporal *Vo*ice *Id*iosyncrasy) corpus [7, 17] was constructed to investigate speaker individualistic temporal characteristics in the speech signal. For the present study, the read speech of the corpus was analysed (16 native speakers of Zürich German × 256 sentences = 4,096 sentences; wav

audio format; sampling frequency = 44.1 kHz; quantisation depth = 16 bits). For more of the corpus construction, please refer to [7, 17].

2.2. Measurements

All the sound files in the corpus were labelled using Praat [3]. Tiers containing on- and off-sets of vocalic and consonantal intervals were employed for the calculations of %V, varcoV, nPVI-V, varcoC, and nPVI-C. Tiers containing on- and off-sets of voiced intervals were used for the calculations of %VO, varcoVO, and nPVI-VO. Tiers containing syllable boundaries as well as syllable peaks were applied to compute varcoPeak, nPVI-Peak, stdevM, varcoM, rPVIm, nPVIm, stdevP, varcoP, rPVIp and nPVIp. Descriptions of all the measures are listed in the Appendix. Praat scripts were applied for the computations, and the results were saved as tab-delineated files before exporting to SPSS [13] for the constructions of neural networks (multilayers perceptron).

2.3. ANN Topologies

The corpus was randomly partitioned into a training set (70% of the corpus) and a test set (30% of the corpus). Three ANNs were modelled based on the same partitioned corpus using *a*. duration metrics only, *b*. intensity metrics only, and *c*. duration cum intensity metrics. The choices of ANN typologies were the same for all three models except the input covariates, which were the duration, intensity and combined metrics respectively. Table 1 presents more details of the ANN architectures, which were configured on a semi-arbitrary basis, because the purpose of the study was to compare the classification strengths rather than maximising classification rates. Nonetheless, we did venture a more complicated configuration of the networks (two hidden layers with 100 neurons in each), but the recognition time increased dramatically without remarkable improvements of the recognition rates.

3. RESULTS AND DISCUSSION

3.1. Speaker Recognition Rates

The average speaker recognition rates yielded from the ANNs in the training set were 17.3% (duration only), 33.1% (intensity only), and 42.3% (duration cum intensity). The mean recognition rates calculated from the test set were 14.2% (duration only), 30.3% (intensity only), and 36.9% (duration cum intensity). Table 2 shows more descriptive statistics of speaker recognition rates in different choices of metrics. Figures 1 and 2 present the breakdowns of classification rates for each speaker in both training and test sets.

Table 1: ANNs fitting information.

Input
Input covariates: duration metrics only; intensity metrics only; duration cum intensity metrics Rescaling method for covariates: Standardised
Hidden layer (1 hidden layer)
Number of neurons in the hidden layer: 10 + 1 bias Activation function: Sigmoid
Output
Dependent variable: speaker Activation function: Softmax Error function: Cross-entropy
---------- NB All networks are feedforward without recursions.

Table 2: Descriptive statistics of speaker recognition rates (in %) with different independent variables.

	mean	std. dev.	std. err.	min.	max.
(i) Training set					
D*	17.3	11.6	2.9	3.7	38.4
I*	33.1	19.5	4.9	10.1	77.6
C*	42.3	17.3	4.3	14.8	77.6
(ii) Test set					
D*	14.2	13.1	3.3	0.0	39.4
I*	30.3	19.2	4.8	9.5	73.2
C*	36.9	19.5	4.9	11.9	67.1

* D = duration metrics; I = intensity metrics; C = duration cum intensity metrics.

Figure 1: Speaker recognition rates in the training set (X-axis: speaker ID; Y-axis: recognition rate in %). The horizontal dashed line indicates the chance level (100% ÷ 16 ≅ 6.3%).

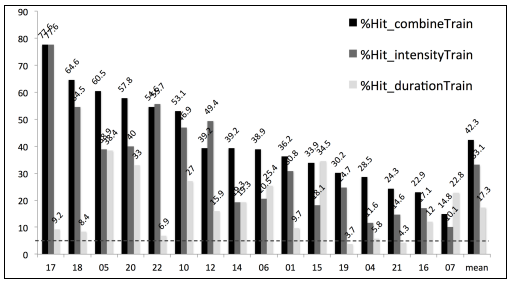
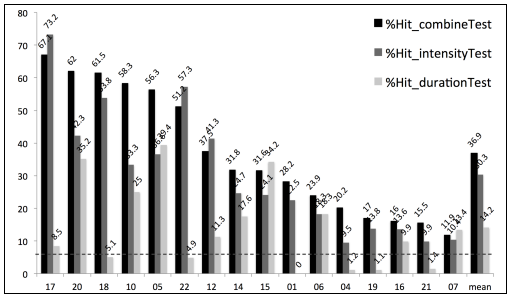


Figure 2: Speaker recognition rates in the test set (X-axis: speaker ID; Y-axis: recognition rate in %). The horizontal dashed line indicates the chance level (100% ÷ 16 ≅ 6.3%).



3.2. Comparisons of Recognition Strengths

First of all, the distribution normalities of the recognition rates from three ANN models (both training and test data) were evaluated using the Shapiro-Wilk test, and the results indicated no serious deviations from normality (all p values ≥ 0.05).

Table 3: Results of Bartlett's tests and paired t-tests (2-sided).

Train vs. Test	Bartlett's tests K^2 (df=1)	p	t-tests t (df=15)	p
duration	0.2107	>0.6	3.7206	=0.002
intensity	0.0056	>0.9	2.0576	>0.05
combined	0.1908	>0.6	3.9097	=0.001

Table 4: Results of Bartlett's tests and ANOVAs.

	Bartlett's tests K^2 (df=2)	p	ANOVAs F (df=2,45)	p
durationTrain intensityTrain combineTrain	3.9451	>0.1	9.4043	<0.0004
durationTest intensityTest combineTest	2.6837	>0.2	7.1583	<0.002

Paired samples t-tests were run in order to compare if the training set recognitions were significantly better than the test set recognitions. Bartlett's tests indicated the data variances were homogenous; therefore, no adjustments were needed. Tables 3 shows the statistical results: only the recognition rates between training and test sets using intensity measures were not significantly different. The results indicated some degrees of over-adaptations of the training data, which is one of the weaknesses of the ANN [16].

Finally, univariate ANOVAs were utilised and the results indicated that significant effects of the metrics choice existed (Table 4 shows the statistics). Bartlett's tests confirmed the equalities of variances, so no adjustments were necessary (also see Table 4).

Post hoc pairwise comparisons (Bonferroni adjusted) indicated that in the training set, intensity metrics and intensity cum duration metrics were significantly better than duration metrics alone at idendifying speakers (${}^{\text{Train}}p_{\text{intensity:duration}} < 0.03$, ${}^{\text{Train}}p_{\text{combine:duration}} < 0.0003$). However, the intensity metrics and the combined metrics were not significantly different (${}^{\text{Train}}p_{\text{intensity:combine}} > 0.4$). The test set showed similar patterns: intensity metrics and combined metrics performed significantly better in speaker recognitions, but the intensity metrics and combined metrics were not significantly different (${}^{\text{Test}}p_{\text{intensity:duration}} < 0.04$, ${}^{\text{Test}}p_{\text{combine:duration}} < 0.002$, ${}^{\text{Test}}p_{\text{intensity:combine}} > 0.8$). Figure 3 visualises the patterns.

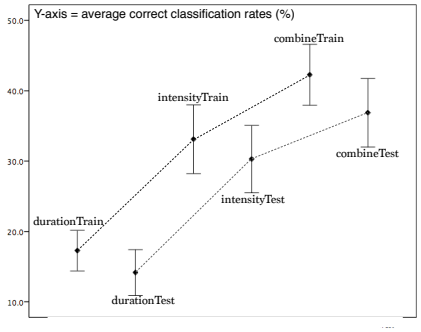
This suggests that although both duration and intensity measures had significant speaker effects [7, 17, 10, 2], intensity variability in the speech signal showed more strength compared with duration metrics to classify speakers. However, albeit the speaker discriminability of duration cum intensity measures were not significantly different from intensity measures in post hoc tests, the significance level of $p_{\text{intensity:duration}}$ (0.03) was a hundred folds the significance level of $p_{\text{combine:duration}}$ (0.0003) in the training set. In the test set, the significance level of $p_{\text{intensity:duration}}$ (0.04) was twenty folds the significance level of $p_{\text{combine:duration}}$ (0.002).

In other words, alghough both combined metrics and intensity metrics were significantly better than duration metrics for the recognition of speakers in both training and test sets, the probability that the combined metrics significantly improved over duration metrics increased 2.97 percentage points than the intensity metrics alone for the training data ($(1 - {}^{\text{Train}}p_{\text{combine:duration}}) - (1 - {}^{\text{Train}}p_{\text{intensity:duration}}) = (1 - 0.0003) - (1 - 0.03) = 0.0297$), and 3.8 percentage points for the test data ($(1 - {}^{\text{Test}}p_{\text{combine:duration}}) - (1 - {}^{\text{Test}}p_{\text{intensity:duration}}) = (1 - 0.002) - (1 - 0.04) = 0.038$).

In addition, it was also assumed that only two rounds of speaker classification tasks (on the basis of intensity cum duration metrics and intensity metrics alone) had been performed. Paired samples t-tests showed that the combined metrics significantly improved recognition rates than intensity metrics alone in the training set (t = 4.1527,

2-sided, $p < 0.0009$, with $df = 15$) and test set ($t = 2.9588$, 2-sided, $p < 0.01$, with $df = 15$).

Figure 3: Error bar graph showing general speaker recognition rates (mean ± 1 standard error) with different metrics choices.



5. CONCLUSION

The present study explored speaker recognition strengths using duration variability, intensity variabily and the two combined in the speech signals with the feedforward ANN. The results suggested that intensity metrics and intensity cum duration metrics were stronger in speaker recognitions.

Compared with our previous studies, we can see that speaker recognition success depends on the recognition algorithms as well. For instance, the TEVOID corpus with the intensity metrics as described in the current study alone yielded different degrees of correct classifications using the *k*-nearest neighbours (*k*NN), feedforward ANN, and multinomial logistic regressions [2, 11], where the *k*NN showed poorest performance (average hit rate ≅ 12%), and the logistic regression showed the best performance (average hit rate ≅ 38%). There is potential to design or optimise recognition algorithms and achieve higher recognition rates with the combination of intensity and duration measures.

In addition, we have also observed that although significant between-speakers variability has been proven by statistical tests, it does not necessarily entail high recognition rates using available classification algorithms (duration metrics in particular).

Open questions for future research are how robust the presented measures are in the context of degraded and distorted speech. Also, how speaker recognition rates would increase if the duration and intensity measures are coupled with spectral measurements is worth further examinations. Moreover, how should the metrics and classification algorithms be optimised is also subjective to further investigations. On the engineering side, a fully automatic extraction of the metrics from the acoustic signal is also in our research agenda.

6. APPENDIX−METRICS DESCRIPTIONS

6.1. Duration Metrics

- %V: Percentage of vocalic interval durations out of the total sentential duration.
- %VO: Percentage of voiced interval durations out of the total sentential duration.
- varcoC: Variation coefficient (standard deviation ÷ mean) of consonantal interval durations.
- nPVI-C: Mean of locally averaged pairwise consonantal interval duration differences.
- varcoV: Variation coefficient of vocalic interval durations.
- nPVI-V: Mean of the locally averaged pairwise vocalic interval duration differences.
- varcoPeak: Variation coefficient of syllabic peak-to-peak interval durations.
- nPVI-Peak: Mean of the locally averaged pairwise syllabic peak-to-peak interval duration differences.
- varcoVO: Variation coefficient of voiced interval durations.
- nPVI-VO: Mean of the locally averaged pairwise voiced interval duration differences.

6.2. Intensity Metrics

- stdevM/P: Standard deviation of mean/peak intensity of each syllable.
- varcoM/P: Variation coefficient of mean/peak intensity of each syllable.
- rPVIm/p: Mean of the pairwise mean/peak intensity differences of consecutive syllables.
- nPVIm/p: Mean of the locally averaged pairwise mean/peak intensity differences of consecutive syllables.

Mathematical formulae of these metrics can be found in [17] and [10, 11].

7. ACKNOWLEDGEMENTS

This study is supported by the Gebert Rüf Stiftung (Grant No. GRS-027/13) and the Swiss National Science Foundation (Grant No. 100015_135287). The authors would like to thank Adrian Leemann and Marie-José Kolly for their work on the TEVOID corpus

8. REFERENCES

[1] Abercrombie, D. 1967. *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

[2] He, L., Dellwo, V. submitted. The role of syllable intensity in between-speaker rhythmic variability.

[3] Boersma, P., Weenink, D. 2014. Praat: doing phonetics by computer (version 5.3.65). http://www.praat.org/.

[4] Classe, A. 1939. *The Rhythm of English Prose* Oxford: Blackwell.

[5] Dellwo, V. 2006. Rhythm and speech rate: A variation coefficient for deltaC. In: Karnowski, P., Szigeti, I. (eds), *Language and Language Processing*, Frankfurt: Peter Lang, 231-241.

[6] Dellwo, V., Huckvale, M., Ashby, M. 2007. How is individuality expressed in voice? An introduction to speech production and description for speaker classification. In: Müller, C., (ed), *Speaker Classification I*. Berlin and Heidelberg: Springer Verlag, 1-20.

[7] Dellwo, V., Leemann, A., Kolly, M.-J. 2012. Speaker idiosyncratic rhythmic features in the speech signal. *Proc. Interspeech 2012* Portland, 1582-1585.

[8] Gershenson, C. 2003. Artificial neural networks for beginners. arXiv: cs/0308031. http://arxiv.org/pdf/cs/0308031.pdf.

[9] Grabe, E., Low, E. L. 2002. Durational variability in speech and rhythm class hypothesis. In: Warner, N., Gussenhoven, C. (eds), *Papers in Laboratory Phonology 7*. Berlin: Mouton de Gruyter, 515-543.

[10] He, L., Dellwo, V. 2014. Speaker idiosyncratic variability of intensity across syllables, *Proc. Interspeech 2014*, Singapore, 233-237.

[11] He, L., Glavitsch, U., Dellwo, V. 2014. Automatic speaker identification using syllable intensity variability: an initial attempt using the *k*NN classifier. Abstract presented at Phonetik & Phonologie 10, Konstanz. http://ling.unikonstanz.de/pages/conferences/pp10/abstracts/He_pp10.pdf.

[12] Heaton, J. 2011. *Introduction to the Math of Neural Networks (Beta-1)*. Chesterfield: Heaton Research Inc. (http://www.heatonresearch.com).

[13] IBM Corp. 2013. IBM SPSS Statistics for Macintosh (version 22.0). Armonk, NY: IBM Corp.

[14] Jain, A. K., Mao, J. 1996. Artificial neural networks: a tutorial. *Computer*. 29, 31-44.

[15] Jassem, W., Grygiel, W. 2004. Off-line classification of Polish vowel spectra using artificial neural networks. *J. IPA*. 34, 37-52.

[16] Krogh, A. 2008. What are artificial neural networks? *Nat. Biotechnol*. 26, 195-197.

[17] Leemann, A., Kolly, M.-J., Dellwo, V. 2014. Speech-individuality in suprasegmental temporal features: implications for forensic voice comparison. *Forensic Sci. Int*. 238, 59-67.

[18] Lloyd James, A. 1940. *Speech Signals in Telephony* London: Sir Isaac Pitman & Sons.

[19] Low, E. L., Grabe, E., Nolan, F. 2000. Quantitative characterization of speech rhythm: Syllable-timing in Singapore English. *Lang. Speech*. 43, 377-401.

[20] Matovski, D. S., Nixon, M. S., Mahmoodi, S., Carter, J. M. 2010. The effect of time on the performance of gait biometrics. In: *Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*. Washington DC. http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5618724.

[21] Pike, K. 1945. *The Intonation of American English*. Ann Arbor: University of Michigan Press.

[22] Ramus, F., Nespor, M., Mehler, J. 1999. Correlates of linguistic rhythm in the speech signal. *Cognition*. 73, 265-292.

[23] Sałapa, K., Trawińska, A., Roterman-Konieczna, I. 2013. Forensic voice comparison by means of artificial neural networks. *Bio-Alg. Med-Syst*. 9, 191-197.

[24] Sałapa, K., Trawińska, A., Roterman, I., Tadeusiewicz, R. 2014. Speaker identification based on artificial neural networks. Case study: the Polish vowel a (pilot study). *Bio-Alg. Med-Syst*. 10, 91-99.

[25] Söder, O. 2008. kNN classifiers 1: What is a kNN classifier? http://www.fon.hum.uva.nl/praat/manual/kNN_classifiers_1__What_is_a_kNN_classifier_.html

[26] Weenink, D. 2006. Speaker-Adaptive Vowel Identification. Doctoral dissertation, Universiteit van Amsterdam.

[27] White, L., Mattys, L. S. 2007. Calibrating rhythm: first language and second language studies. *J. Phonet*. 35, 501-522.

[28] Yoo, J.-H., Hwang, D., Moon, K.-Y., Nixon, M. S. 2008. Automated human recognition by gait using neural network. In: *First Workshops on Image Processing Theory, Tools and Applications*. Sousse. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4743792.

■

Study III

The role of syllable intensity in between- speaker rhythmic variability

This work has been published in December 2016 in the *International Journal of Speech, Language and the Law* (Volume 23, Issue 2, pp. 243-273).
Official URL (DOI):
http://dx.doi.org/10.1558/ijsll.v23i2.30345

The International
Journal of
Speech,
Language
and the Law

IJSLL (PRINT) ISSN 1748-8885
IJSLL (ONLINE) ISSN 1748-8893

Article

The role of syllable intensity in between-speaker rhythmic variability

Lei He and Volker Dellwo

Abstract

Speech rhythm in terms of durational variability of different levels of phonetic intervals can vary between speakers. The present article examines the role of syllabic intensity characteristics in rhythmic variability. Mean and peak intensity variability across syllables (stdevM, varcoM, stdevP, varcoP, rPVIm, nPVIm, rPVIp, nPVIp; henceforth: intensity measures) were investigated as a function of speaker in a database where within-speaker variability was strong (BonnTempo) and another database designed to examine between-speaker rhythmic variability (TEVOID). It was found that the intensity measures varied significantly between speakers in both databases. Semiautomatic speaker recognition based on duration measures (%V, ΔV(ln), ΔC(ln), ΔPeak(ln), ΔSyll(ln) and nPVISyll) and intensity measures using multinomial logistic regression and feedforward neural networks was carried out for the two databases. Results showed that intensity measures contained stronger speaker specific information compared to measures based on durational variability of phonetic intervals. In addition, effects of the recognition algorithms (speaker recognition using multinomial logistic regression was significantly better than neural networks for BonnTempo) and data normalisation procedures (z-score normalised data was significantly better than non-normalised data in TEVOID) were discovered. This means that syllable intensity characteristics play an important role in between-speaker rhythmic differences and possibly in speech rhythm variability in general.

KEYWORDS SPEECH RHYTHM, SPEECH INTENSITY, ARTICULATION, SPEAKER-IDIOSYNCRATIC FEATURES

Affiliation
University of Zurich
email: lei.he@uzh.ch volker.dellwo@uzh.ch

IJSLL VOL 23.2 2016 243–273 doi : 10.1558/ijsll.v23i2.30345

1. Introduction

While it is very plausible that the durations of phonetic intervals like segments or syllables contribute to speech rhythm, hardly any attention has been paid to the fact that the rhythmicity of a signal created by a sequence of events not only depends on the durational relationships between these events, but also on their intensity differences. This article serves to investigate speech rhythm in terms of such intensity differences in the context of differences between individual speakers. In the following paragraphs, we review literature from the perspectives of a) models of speech rhythm, indicating a lacuna of intensity based research (Section 1.1), and b) the rationale for examining the idiosyncratic rhythm of speakers, especially from the angle of intensity variability (Section 1.2).

1.1. The scope of research on speech rhythm and associated models of rhythm

There have been a large number of studies on speech rhythm variability, focusing on different aspects of speech: between-language rhythmic similarities and differences (e.g., Abercrombie 1967; Grabe and Low 2002; Loukina, Kochanski, Rosner, Keane and Shih 2011; Ramus, Nespor and Mehler 1999; Tilsen and Arvaniti 2013), rhythmic characteristics of dialects or vernaculars of a language (e.g., Frota and Vigàrio 2001; Low, Grabe and Nolan 2000; Rathcke and Smith 2015; White, Payne and Mattys 2009), metrically regular speech (e.g., Leong, Stone, Turner and Goswami 2014; O'Dell and Nieminen 1999), child and child-directed speech (e.g., Lee, Kitamura, Burnham and Todd 2014; Payne, Post, Astruc, Prieto and Vanrell 2012; Polyansakaya and Ordin 2015), pathological speech (e.g., Leong and Goswami 2014; Liss, White, Mattys, Lansford, Lotto, Spitzer and Caviness 2009; White, Liss and Dellwo 2010), and speaker idiosyncratic rhythmic characteristics (e.g., Dellwo, Leemann and Kolly 2012, 2015; Leemann, Kolly and Dellwo 2014).

How is speech rhythm defined in these studies? A consensual definition is yet to be reached, but two aspects, timing and spectral prominence, are plausible mechanisms underpinning the phenomenon (Nolan and Jeon 2015). It is interesting, however, that nearly all models of rhythm are heavily based on durational characteristics of the speech signal in some way. Early speculation focused on the durations of syllables and feet (Abercrombie 1967). More recently, the durations of vocalic and intervocalic (consonantal) intervals have been demonstrated to be more perceptually salient in terms of speech rhythm, whence the genesis of the widely used rhythm metrics (e.g., Ramus, Nespor and Mehler 1999; Grabe and Low 2002; Dellwo 2006, 2009). Another approach to speech rhythm, the coupled-oscillator model, measures the phase relationships between varying recurring perceptual beats at different levels (e.g., syllable level and stress level) in the speech signal (Barbosa 2002; Cummins and Port 1998; O'Dell and Nieminen

1999). But again, the phase relationship between these levels is only another way of expressing durational differences between perceptually salient intervals of the signal. Similar to this model, the amplitude modulation phase model describes the phase relationships between stress rate and syllable rate amplitude modulations, which underpin the perceived rhythmic patterns (Leong et al. 2014). Other models of speech rhythm include the auditory primal sketch model that infers temporal patterns of syllabic or sonorant events together with their prominences derived from duration, frequency and intensity information (Lee and Todd 2004; Lee et al. 2014), and the low-frequency Fourier analysis of the amplitude envelope (Tilsen and Johnson 2008; Tilsen and Arvaniti 2013) which showed the durational regularities of recurring low-frequency spectral prominences in speech. Although these two approaches differ from the previous ones in that they do not directly measure duration in terms of interval onsets and offsets, they all search for durationally recurring patterns in the time domain. Durational variability, however, is not the only factor contributing to the perceived rhythmicity of an utterance. The perceptual presence of a rhythmic unit is also marked by its relative intensity (Kohler 2008, 2009). Nevertheless, speech rhythm models focusing on the intensity variability in the signal are only sporadic: Low (1998) on the rhythmic characteristics of Singapore English; He (2012) on the rhythmic differences between first and second language English, and Cichocki, Selouani and Perreault (2014) on the rhythmic differences between Canadian French dialects. Other models that go beyond temporal characteristics include pitch-integrated models (Cumming 2011) and loudness-integrated models (Fuchs 2014; Gavles, Garcia, Duarte and Galves 2002).

In summary, models of speech rhythm are dominated by the search for temporal patterns in the signal. This is also true for studies analysing the variability of speech rhythm between speakers (Arvaniti 2012; Dellwo, Leemann and Kolly 2012, 2015; Leemann, Kolly and Dellwo 2014; Shriberg, Ferrer, Kajarekar, Venkataraman and Stolcke 2005; Wiget, White, Schuppler, Grenon, Rauch and Mattys 2010; Yoon 2010). In the upcoming section, the rationale for studying between-speaker rhythm differences, especially from the perspective of intensity variability, is introduced.

1.2. Why should intensity play a role in rhythmic variability between speakers?

A possible rationale motivating rhythmic variability between speakers was derived from the observation that the kinematic properties of the articulators over time are, on the one hand, driven by their individual anatomic characteristics, e.g. their spatial dimensions, mass and accelerations (Perrier 2012), and, on the other hand, by the individual ways speakers acquired to operate their articulators (Dellwo, Leemann and Kolly 2015; Wretling and Eriksson 1998). The individual steering of the articulators should then result in individual temporal

characteristics of speech. Dellwo, Leemann and Kolly (2015), Leemann, Kolly and Dellwo (2014) and Wiget et al. (2010) showed that durational measures of speech rhythm could vary strongly and significantly between speakers. Dellwo, Leemann and Kolly (2015) further revealed that the most likely sources of this variability are articulatory factors varying between speakers.

In terms of speaker-individual differences, it seems that a possible contribution of the individual kinematics of the articulators which might lead to individual temporal characteristics in the signal may also influence aspects of the signal that stand in relation to its intensity. In fact, there is evidence that the degree of mouth aperture is related to the intensity of the speech signal (Birkholz, Kröger and Neuschaefer-Rube 2011; Erickson, Kim, Kawahara, Wilson, Menezes, Suemitsu and Moore 2015; Garnier, Wolfe, Henrich and Smith 2008; and notably Chandrasekaran, Trubanova, Stillittano, Caplier and Ghazanfar 2009), and that subglottic air pressure and speech intensity are related in complex fashion in individuals (Plant and Younger 2000). Individual differences in the movements as well as the anatomic characteristics of the organs of speech should thus also give rise to between-speaker intensity variability. Another contribution to individual intensity contour characteristics is pulmonic air pressure, the energy source that is universal to all speech production. Wilson and Leeper (1992) found in a well-controlled syllable context that subglottal pressures ranged from 4.5 to 12.8 cm H_2O for men and from 3.8 to 12.6 cm H_2O for women with normal speech effort. In connected speech, individual differences in subglottal air pressure changes could be even more conspicuous. This could inevitably result in idiosyncratic intensity fluctuations in the speech signal. In He and Dellwo (2014), support for this assumption has already been found by demonstrating that significant speaker effects of intensity variability exist, both in terms of mean intensity and peak intensity variability across syllables.

The knowledge about speaker-specific rhythm can potentially be applied in forensic phonetic caseworks. Typically, the expert would estimate the likelihood of whether two or more speech samples are from the same or different speakers using a suite of acoustic measures (Dellwo 2015). However, the trace materials are often degraded in various ways, such as telephone line transmission and voice disguise. Both affect acoustic characteristics such as fundamental and resonance frequencies (e.g., Byrne and Foulkes 2004; Eriksson and Wretling 1997; Jovičić, Jovanović, Subotić and Grozdić 2015). However, durational rhythm measures are not much affected either by mobile phone transmission (Leemann, Kolly and Dellwo 2014) or voice disguise (Leemann and Kolly 2015). Although the ways telephony and audio compression would affect intensity measures are still subject to further research (see discussion), they are potentially useful to forensic experts as well because intensity fluctuations may not be a strategy for voice dis-

guise. Licklider and Pollack (1948) and later Kolly and Dellwo (2014) showed that infinitely peak-clipped speech (intensity variability is absent in the signal) is highly intelligible, suggesting that intensity variability in broadband signals has comparatively small auditory effects on the speech signal. Hence, changing intensity patterns may not be a disguise strategy because of a lack of possible auditory feedback.

1.3. The present study

In the present study we carried out an in-depth analysis of between-syllable intensity variability as a source of between-speaker rhythmic differences. The intensity variability was quantified based on the degrees of dispersion of syllable intensity levels (both average and peak values), and the mean differences between consecutive syllable intensity levels (both average and peak values). We measured the intensity variability of a sentence in terms of the standard deviation of peak (stdevP) and mean (stdevM) syllabic intensity. Intensity, however, is a characteristic of speech that is possibly easy to distort, as only a turn of the head can lead to a drastic drop of the overall intensity at the receiver's ear (or a microphone). The standard deviation of intensity measured over an utterance may be affected by such artefacts. For this reason, we also measured the proportional intensity differences between consecutive syllables, which are possibly not affected by such changes, or only marginally so. We therefore used a measure that has originally been applied to measuring the average difference between consecutive consonantal or vocalic interval durations (Pairwise Variability Index, henceforth PVI, Grabe and Low 2002) and measured the average mean (nPVIm) and peak (nPVIp) intensity differences between consecutive syllables (see Section 2.2 as well as He and Dellwo 2014 for details). To make our studies comparable, we used the same databases as Dellwo, Leemann and Kolly (2015), the BonnTempo and the TEVOID corpora. We addressed the following points:

a. We investigated between-speaker variability to find the intensity measures that best account for this variability (Section 3.2) using multinomial logistic regression (see Leemann, Kolly and Dellwo 2014 for a similar test on duration-based rhythm measures). Including measures which contribute less to between-speaker variability may confuse the recognition algorithms that we would use and thus result in poorer speaker-recognition performance.

b. We trained a semiautomatic speaker-recognition[1] model to compare the performance of the model for intensity and duration variability measures as described in Dellwo, Leemann and Kolly (2015) (Section 3.3). With this we aimed at retrieving some information about which of the dimensions contains more speaker-specific information. More specifically, we carried out the following analyses:

1. We compared duration measures with intensity measures and the effect of combining the two to address the question of which domain of measures provided higher recognition results.
2. We compared the effects of two different recognition algorithms: multinomial logistic regression and feedforward neural networks. We chose these two algorithms because they differ in terms of working mechanisms. Multinomial logistic regression is based on logit-probability and feedforward neural networks, on connectionist computing. Algorithms with similar mechanisms, such as the neural networks and support vector machine, may yield very similar results, thus were avoided. We were interested in the role of the algorithm in the speaker-recognition performance based on our measures.
3. We tested the effect of z-score normalisation by sentence on speaker-recognition performance. A closed-set text-dependent speaker-recognition system usually has a known sentence bank. We investigated whether normalising for variability induced by sentences can increase speaker-recognition performance.

We expect that our performance for semiautomatic speaker recognition based on intensity and duration information should be low in comparison to state-of-the-art speaker-recognition systems, as only single dimensional information is used. With this experiment, however, we aimed at interpreting the relative importance of the two dimensions, instead of developing ready-to-use speaker-recognition engines.

2. Method

2.1 The databases

Two databases, BonnTempo and TEVOID, were used in this study for the following reasons: 1) It was possible to compare results from a previously published study (Dellwo, Leemann and Kolly 2015) where the same databases were used; 2) within-speaker variability was strong in BonnTempo. It would be desirable if between-speaker effect was significant despite strong within-speaker variability. Moreover, TEVOID was specifically designed to examine between-speaker rhythmic variability; 3) both databases include German speakers. Although the two databases differ in different dialects of German (BonnTempo: northern German; TEVOID: Zürich German), we do not envision a large between-dialect rhythmic difference.

2.1.1 The BonnTempo database

The BonnTempo database (Dellwo, Steiner, Aschenberner, Dankovičová and

Wagner 2004; Dellwo, Leemann and Kolly 2015; Dellwo 2010) was built for examining speech rhythm in relation to speech rate, hence results in high within-speaker variability. Only German speech data were analysed. The speakers (n = 12, 5 males and 7 females) were recorded in an anechoic chamber with a large membrane condenser microphone (sampling frequency = 44.1 kHz; quantisation depth = 16 bits; WAV format) so that reliable RMS measurements were warranted. Each speaker read seven sentences (number of syllables = 76) at five subjective tempo versions: normal, slow (slow1), slower (slow2), fast (fast1), fastest possible (fast2). More details about the database and data elicitation are available in Dellwo (2010) and Dellwo, Leemann and Kolly (2015). Annotations of syllable onsets and offsets were based on phonological criteria unless relevant acoustic traces were absent due to elision (Dellwo 2010). Appendix A contains the reading material for this database.

2.1.2 The TEVOID database

The TEVOID (*Te*mporal *Vo*ice *Id*iosyncrasy) database (Dellwo, Leemann and Kolly 2012, 2015; Leemann, Kolly and Dellwo 2014) was constructed to investigate individual characteristics of rhythmic variability in the speech signal. Sixteen native speakers of Zürich German (8 males and 8 females) were recorded in a sound-treated booth, each producing 256 read sentences and 16 spontaneous sentences (sampling frequency = 44.1 kHz; quantisation depth = 16 bits; WAV format). For the present study, only read sentences (256 sentences × 16 speakers = 4,096 sentences) were analysed. Details of data elicitation procedures and hardware are available in Leemann, Kolly and Dellwo (2014) and Dellwo, Leemann and Kolly (2012, 2015). Syllable boundaries were annotated automatically based on sound sonority rules; sonority scales were manually attributed to each segment type (Leemann, Kolly and Dellwo 2014). Appendix A shows the first 15 sentences used in this database.

2.2 Signal processing and measurements

We used Praat (Boersma and Weenink 2014) for signal processing and measurements. To calculate an intensity contour we created a 'Praat intensity object' of a speech signal which includes the following processing steps: first the mean amplitude of all sample values is subtracted from the signal (DC bias removal), and then all sample values are squared. A Kaiser window (window coefficient β = 20, sidelobe attenuation $\simeq -190$ dB) with the length of 32 ms is multiplied repeatedly with the squared signal with a window forward of a quarter of the window duration (8 ms), leading to a 75% overlap between windows. For each windowed frame, the sum of squares (SS) of the sample values is calculated and plugged in the formula $10 \times \log_{10}\{[SS/(2\times10^{-5})^2]/0.032\}$ to obtain the intensity level (unit: dB re 20 μPa) in this particular frame.

From the intensity contour we measured both mean (M) and peak (P) intensity of the part of the signal that corresponded to a speech syllable. Mean syllable intensity was calculated as the sum of intensity values in a contour between syllable onset and offset divided by the syllable duration. Peak intensity was measured at the syllable peak point interpolated with the cubic function (see Boersma 1998 for the formulae). The peak point was derived from the amplitude envelope extracted by low-pass filtering the full-wave rectified speech signal at 10 Hz. For both mean and peak intensity we carried out the following measurements (see Table 1(i) for the formulae):

- The standard deviation of mean (stdevM) and peak (stdevP) syllable intensity for each sentence utterance in both databases (BonnTempo and TEVOID). We normalised stdevM and stdevP by taking their variation coefficients (100 × standard deviation / mean, henceforth varcoM and varcoP). This was done to normalise for the fact that some speakers spoke louder than others, or were recorded with higher gains.
- The Pairwise Variability Index of mean (PVIm) and peak (PVIp) syllable intensity, again for each sentence utterance in both databases. It was calculated by taking the average of intensity differences between consecutive syllables in an utterance. PVIm and PVIp were normalised by dividing each pairwise mean or peak intensity difference by their local mean, hence nPVIm and nPVIp. The non-normalised measures are henceforth referred to as 'raw' measures (rPVIm and rPVIp).

Additionally, the initial and final syllables of the sentences were excluded from analysis because, in some cases, the durations of these syllables were too short for the analysis window, and the peak intensity values could not be calculated.

The durational rhythm measures (%V, ΔV(ln), ΔC(ln) and ΔPeak(ln)) were taken from Dellwo, Leemann and Kolly (2015; see Table 1(ii.a) for the formulae). In addition, two other measures of syllable duration (standard deviation of natural logarithm normalised syllable duration and normalised pairwise variability index of syllable duration, i.e., ΔSyll(ln) and nPVISyll; see Table 1 (ii.b) for the formulae) were also tested in terms of their contributions to speaker recognition, because we are primarily interested in syllable-sized intensity variability, and it would be interesting to test whether syllable-sized duration variability plays a role at all.

2.3 Statistical analyses

To test the significance of between-speaker variability on the intensity measures in the BonnTempo and TEVOID databases (Section 3.2), mixed-effects models (fitted by maximum likelihood) were employed using the R package *lme4* (Bates, Maechler, Bolker and Walker 2014). Prior to mixed-effects model analyses, corre-

Table 1: The formulae for calculating intensity and duration variability.

(i) Intensity measures

a. Raw measures

	Holistic measures[a]	Local measures[b]
Mean intensity variability	$\text{stdevM} = \sqrt{\frac{\sum_{j=1}^{n} L_{\text{M}j}^2 - \left[\sum_{j=1}^{n} L_{\text{M}j}^2\right]/n}{n-1}}$	$\text{rPVIm} = \sum_{j=1}^{n-1} \left\lvert L_{\text{M}j} - L_{\text{M}j+1} \right\rvert /(n-1)$
Peak intensity variability	$\text{stdevP} = \sqrt{\frac{\sum_{j=1}^{n} L_{\text{P}j}^2 - \left[\sum_{j=1}^{n} L_{\text{P}j}^2\right]/n}{n-1}}$	$\text{rPVIp} = \sum_{j=1}^{n-1} \left\lvert L_{\text{P}j} - L_{\text{P}j+1} \right\rvert /(n-1)$

b. Normalised measures

	Holistic measures	Local measures
Mean intensity variability	$\text{varcoM} = \frac{\text{stdevM}}{\overline{L_{\text{M}}}} \times 100$	$\text{nPVIm} = \sum_{j=1}^{n-1} \left\lvert \frac{L_{\text{M}j} - L_{\text{M}j+1}}{\left[L_{\text{M}j} + L_{\text{M}j+1}\right]/2} \right\rvert \times \frac{100}{n-1}$
Peak intensity variability	$\text{varcoP} = \frac{\text{stdevP}}{\overline{L_{\text{P}}}} \times 100$	$\text{nPVIp} = \sum_{j=1}^{n-1} \left\lvert \frac{L_{\text{P}j} - L_{\text{P}j+1}}{\left[L_{\text{P}j} + L_{\text{P}j+1}\right]/2} \right\rvert \times \frac{100}{n-1}$

L_{Mj} = the mean intensity of the jth syllable; L_{Pj} = the peak intensity of the jth syllable; $\overline{L}_{M}$ = the average of L_{Mj} in a sentence; $\overline{L}_{P}$ = the average of L_{Pj} in a sentence; n = the number of syllables to be calculated in a sentence.

(ii) Duration measures

a. Vocalic, consonantal and inter-peak interval duration (Dellwo, Leemann and Kolly 2015)

$$\%\text{V} = \frac{\sum_{i=1}^{n_{\text{V}}} V_i}{\sum_{i=1}^{n_{\text{V}}} V_i + \sum_{i=1}^{n_{\text{C}}} C_i} \times 100\%$$

n_V = the number of vocalic intervals in a sentence; n_C = the number of consonantal intervals in a sentence; V_i = duration of the ith vocalic interval; C_i = duration of the ith consonantal interval.

$$\Delta\text{Invl(ln)} = \sqrt{\frac{n_{\text{Invl}} \sum_{i=1}^{n_{\text{Invl}}} (\ln Invl_i)^2 - \left[\sum_{i=1}^{n_{\text{Invl}}} (\ln Invl_i)\right]^2}{n_{\text{Invl}}(n_{\text{Invl}} - 1)}}$$

Invl = interval under observation, either vocalic (V), consonantal (C), or inter-peak[c] (Peak); n_{Invl} = the number of respective intervals in a sentence; $Invl_i$ = duration of the ith interval.

b. Syllable duration

$$\Delta\text{Syll(ln)} = \sqrt{\frac{n \sum_{i=1}^{n} (\ln Syll_i)^2 - \left[\sum_{i=1}^{n} (\ln Syll_i)\right]^2}{n(n-1)}} \qquad \text{nPVISyll} = \sum_{i=1}^{n-1} \left\lvert \frac{Syll_i - Syll_{i+1}}{\left[Syll_i + Syll_{i+1}\right]/2} \right\rvert \times \frac{100}{n-1}$$

$Syll_i$ = the duration of the ith syllable; n = the number of syllables in a sentences.

a. The stdev and varco measures capture the overall intensity dispersions in the signal, hence are collectively referred to as holistic measures.
b. The PVI measures capture intensity differences between consecutive syllables, hence are referred to as local measures.
c. Here 'peak' refers to the time points where the syllabic amplitude envelope maxima occur.

lations among all intensity measures were evaluated using R (R Core Team 2014) to exclude measures that highly predict each other (Section 3.1). Tempo (only in BonnTempo) was modelled as a fixed factor; speaker and sentence (in both BonnTempo and TEVOID) were modelled as random intercepts (rationale: speakers were a sample of all the German-speaking population, and sentences were a sample of an infinitely large population of possible German sentences; Baayen 2008). To test the significance of an effect, a reduced model was formed by excluding the effect (fixed or random) in question, and a likelihood ratio test was run between the full and the reduced models. Moreover, multinomial logistic regression models were fitted on both BonnTempo and TEVOID data using SPSS (IBM Corp. 2013) to analyse how much between-speaker variability each intensity measure can explain (Section 3.2). Speaker was modelled as the nominal response variable, and the intensity measures were modelled as the predicting covariates. Relative importance of each intensity measure was defined as the likelihood ratio χ^2 of each measure divided by the sum of the χ^2s of all measures.

To test which set of rhythm measures (intensity- or duration-based) perform better in semiautomatic speaker-recognition experiments, two different detection algorithms were applied: multinomial logistic regression and feedforward neural networks (fitted with SPSS; see Table 2 for the architectures of the neural networks) (Section 3.3). For each algorithm per database, we ran six rounds of speaker-recognition experiments, using non-normalised and z-score normalised intensity measures, duration measures and a combination of both as predictor variables. A z-score normalised measure z_j was calculated as $z_j = (y_j - \bar{y}_j)/\sigma_j$, where y_j = the score of sentence j, $\bar{y}_j$ = the mean, and σ_j = the standard deviation of all y_j.

Table 2: The architectures of the feedforward neural networks for speaker recognition in BonnTempo and TEVOID databases

(i)	Data partition	
	Training set: 70% randomly selected from the complete BonnTempo or TEVOID datasets	
	Test set: The complement sets of the training sets	
(ii)	Input layer	
	Covariates:	Round 1: intensity-based rhythm measures
		Round 2: z-score transformed intensity measures
		Round 3: duration-based rhythm measures
		Round 4: z-score transformed duration measures
		Round 5: covariates in Round 1 and Round 3
		Round 6: covariates in Round 2 and Round 4
	Rescaling method for covariates: Standardised	

(iii)	Hidden layer[a]	
	Number of hidden layers:	1
	Number of neurons in the hidden layer:	10 neurons + 1 bias neuron
	Activation function:	Sigmoid
(iv)	Output layer	
	Output classes:	Speakers (16 for TEVOID; 12 for BonnTempo)
	Activation function:	Softmax
	Error function:	Cross-entropy

a. We piloted different network configurations, the most complicated being two layers with 200 neurons in each, but the results did not differ much from our current configurations.

Finally, to address the questions of 1) which domain of measures (intensity measures, duration measures or a combination of the two) provide high recognition results, 2) whether the choice of detection algorithms (multinomial logistic regression or neural networks) has an effect, and 3) whether an effect of data normalisation (non-normalised vs. z-score normalised recognition rates) exists, we performed 3 (duration vs. intensity vs. combined measures) × 2 (multinomial logistic regression vs. neural networks) × 2 (non-normalised vs. z-score normalised) factorial analyses of variances (ANOVAs) with Type III sum of squares on the recognition results for both BonnTempo and TEVOID databases (Section 3.3).

To test the potential contributions of syllable duration measures, the same speaker-recognition algorithms were used either with only ΔSyll(ln) and nPVISyll, or a combination of all duration measures for both databases.

3. Data analyses and results

3.1 Correlations between intensity measures

Pairwise correlations between the eight intensity measures can be found in Figure 1. Since the numbers of items were large, we applied the rule of $|r| \geq 2/n^{½}$ to determine whether a correlation indicates a relationship (Krehbiel 2004). For BonnTempo, the formula yielded a threshold of 0.338 (n = 7 sentences × 5 tempi). For TEVOID, the formula yielded a threshold of 0.125 (n = 256 sentences). As Figure 1 shows, the correlations of measures within each of the measurement types (mean and peak) were high, whereas the correlations of measures between the measurement types were low. This suggests that, on the one hand, the differ-

ent variants of the mean or peak measures were at least moderately predictable among each other. On the other hand, mean and peak measures should carry different information about intensity variability. Based on this result, we selected only nPVIm and nPVIp for the analyses of speaker effect in the BonnTempo and TEVOID datasets in Section 3.2.

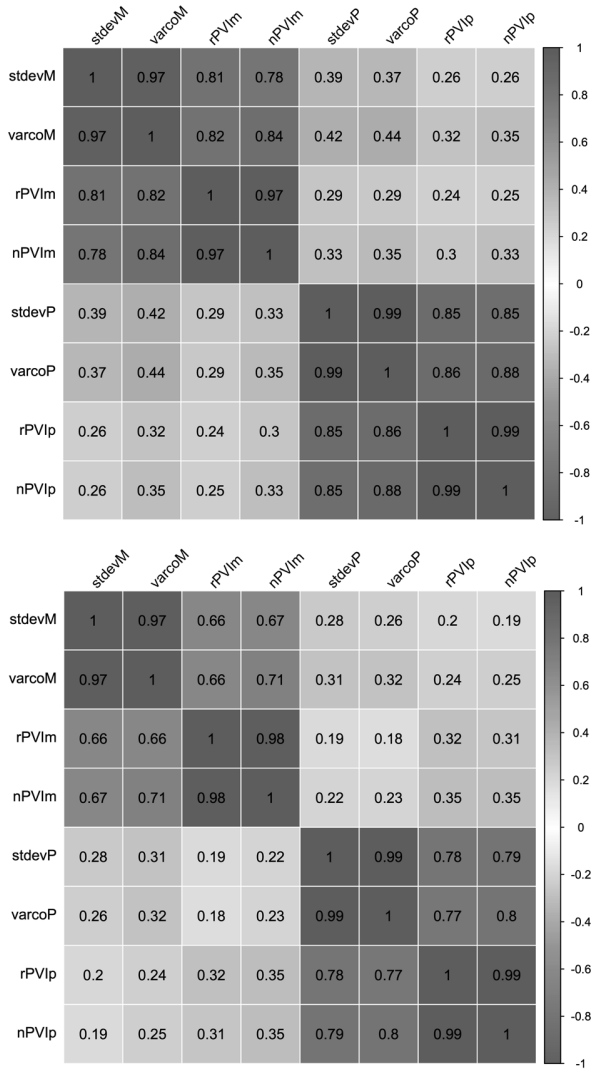


Figure 1: Correlation matrices showing Pearson's correlation coefficient r of the intensity measures in the TEVOID database (top) and the BonnTempo database (bottom). All correlations were highly significant ($p < 0.005$)

3.2 Between-speaker variability and intensity measures

Table 4 presents the results of the mixed-effects models (fitted by maximum likelihood) for between-speaker variability of the intensity measures in the BonnTempo and TEVOID databases. Figure 2 illustrates the between-speaker differences across tempo versions on nPVIm and nPVIp for BonnTempo speakers. Figure 3 illustrates the between-speaker differences on nPVIm and nPVIp for TEVOID speakers. Details of fitted models and R codes are presented in Table 3 and Appendix B.

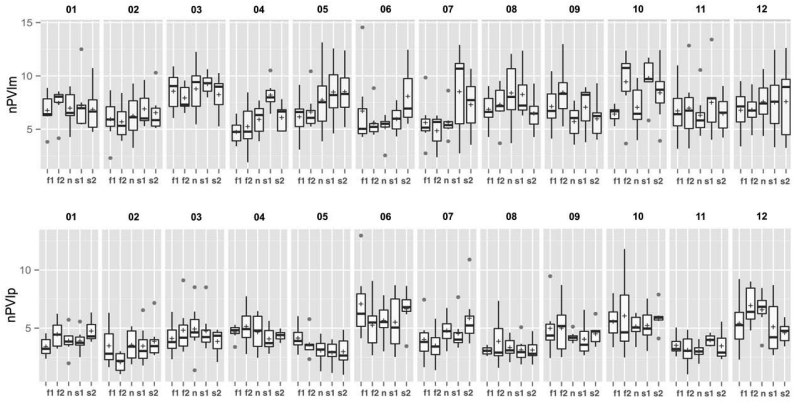


Figure 2: Boxplots showing the distributions of nPVIm and nPVIp of each BonnTempo speaker across five tempo versions (f1 = fast1, f2 = fast2, n = normal, s1 = slow1, s2 = slow2). The following outliers were outside the plotted range: speaker 04 (nPVIm, slow1, sentence 4, value = 16.56), speaker 12 (nPVIm, fast1, sentence 1, value = 19.80; nPVIp, fast1, sentence 1, value = 19.56)

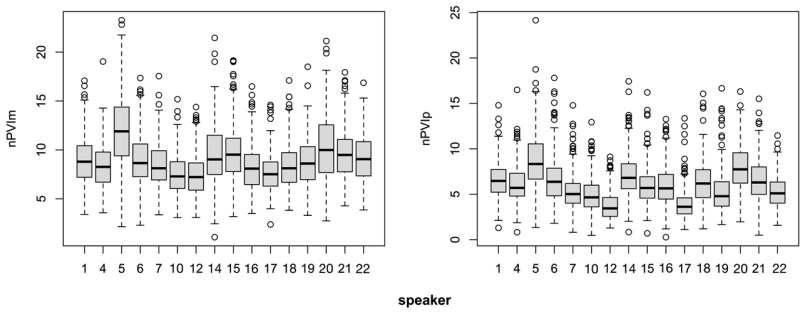


Figure 3: Boxplots showing the distributions of nPVIm and nPVIp of each TEVOID speaker

Table 3: Descriptions of all fitted mixed-effects models with nPVIm and nPVIp as dependent variables for the BonnTempo and TEVOID data

Model ID	Model descriptions			Remarks
	Dependent variable	Fixed effect	Random intercept(s)	
(i) BonnTempo database				
MDL1	nPVIm	tempo	speaker; sentence	full model
MDL2	nPVIm	tempo	sentence	speaker-reduced model
MDL3	nPVIm	1[a]	speaker; sentence	tempo-reduced model
MDL11	nPVIp	tempo	speaker; sentence	full model
MDL22	nPVIp	tempo	sentence	speaker-reduced model
MDL33	nPVIp	1[a]	speaker; sentence	tempo-reduced model
(ii) TEVOID database				
MDL4	nPVIm	1[a]	speaker; sentence	full model
MDL5	nPVIm	1[a]	sentence	speaker-reduced model
MDL44	nPVIp	1[a]	speaker; sentence	full model
MDL55	nPVIp	1[a]	sentence	speaker-reduced model

a. '1' implies that no fixed effect is fitted in the model.

As shown in Table 4, full models are significantly different from speaker-reduced models with increased goodness of fit (smaller AIC values), indicating that between-speaker variation was significant for both nPVIm and nPVIp. In like manner, the tempo effect was tested to be significant for nPVIm, but not for nPVIp. Least-squares means (Tukey adjusted for p values) were compared between tempo levels on nPVIm using lsmeans R package (Lenth 2014). The following pairs were found significantly different: fast1 < slow1 (adjusted $p < 0.0005$), fast2 < slow1 (adjusted $p < 0.004$), and normal < slow1 (adjusted $p < 0.002$).

Table 4: Results of mixed model comparisons of BonnTempo and TEVOID data using likelihood ratio tests. Akaike information criterion values in boldface indicate better fit

Model Comparison	Akaike Information Criterion (AIC)		χ^2 [df]	p[a]
(i) BonnTempo results				
MDL1, MDL2	**1839.9** $_{MDL1}$	1862.5 $_{MDL2}$	24.61[1]	$½\times7.02\times10^{-7}$
MDL1, MDL3	**1839.9** $_{MDL1}$	1854.2 $_{MDL3}$	22.283[4]	< 0.0002
MDL11, MDL22	**1674.0** $_{MDL11}$	1759.4 $_{MDL22}$	87.395[1]	$½\times2.2\times10^{-16}$
MDL11, MDL33	1674.0 $_{MDL11}$	**1668.7** $_{MDL33}$	2.6426[4]	> 0.6

(ii) TEVOID results				
MDL4, MDL5	**17066** $_{MDL4}$	18407 $_{MDL5}$	1343[1]	½×2.2×10^{-16}
MDL44, MDL55	**17565** $_{MDL44}$	18887 $_{MDL55}$	1330[1]	½×2.2×10^{-16}

a. Statistical tests for linear models assume that estimated parameters could have either positive or negative intercepts (or slopes). However, for random effect variances, only positive values are possible, resulting in conservative hypothesis testing. To test the significance of a single random effect, the p-value should be adjusted by halving it (Bolker 2015). That is why the p-values for random effects were multiplied by ½ in this table.

Table 5 shows the results of the multinomial logistic regression to test which of the intensity measures explains the between-speaker variability best. Figure 4 visualises the contribution of each intensity measure in explaining between-speaker variability (see last column of Table 5). For the TEVOID databases, the measures having the highest contributions are varcoP, stdevP, varcoM, and stdevM. For BonnTempo these measures also contribute highly to between-speaker variability even though the difference to the remaining measures is much smaller.

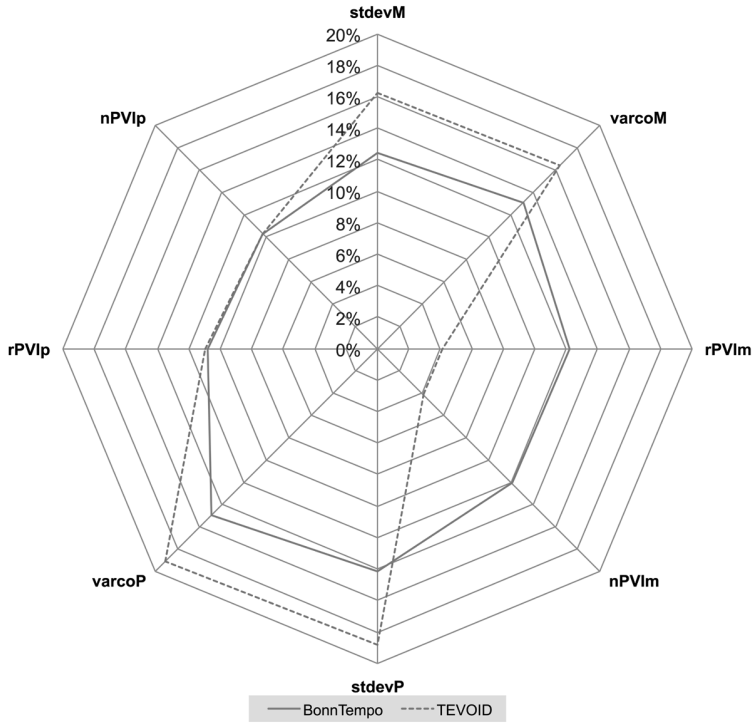


Figure 4: Radar chart illustrating the relative importance of each intensity measure in the multinomial logistic models with speaker as the nominal response variable in the BonnTempo (solid lines; see Table 5(i)) and TEVOID (dotted lines; see Table 5(ii)) databases. The radii of the intensity measures are proportional to their corresponding contributions to explaining the variability between speakers

Table 5: Results of multinomial logistic regressions for both BonnTempo and TEVOID databases

	−2LL	χ^2 [df][a]		p	Variability explained[b]
(i) BonnTempo results					
Model fitting information					
null model	2082.325				
full model	1236.525	845.800[88]		<0.0001	
Likelihood ratio test of each intensity measure					
stdevM	1280.076	43.550 [11]		<0.0001	12.4%
varcoM	1282.339	45.814 [11]		<0.0001	13.1%
rPVIm	1279.255	42.729 [11]		<0.0001	12.2%
nPVIm	1278.855	42.329 [11]	$\Sigma\chi^2$ = 349.928	<0.0001	12.1%
stdevP	1286.105	49.580 [11]		<0.0001	14.2%
varcoP	1288.741	52.216 [11]		<0.0001	14.9%
rPVIp	1274.176	37.651 [11]		<0.0001	10.8%
nPVIp	1272.584	36.059 [11]		<0.0001	10.3%
(ii) TEVOID results					
Model fitting information					
null model	22713.047				
full model	15283.135	7429.912[120]		<0.0001	
Likelihood ratio test of each intensity measure					
stdevM	15602.302	319.167 [15]		<0.0001	16.2%
varcoM	15606.302	323.167 [15]		<0.0001	16.4%
rPVIm	15363.758	80.623 [15]		<0.0001	4.1%
nPVIm	15364.145	81.010 [15]	$\Sigma\chi^2$ = 1967.634	<0.0001	4.1%
stdevP	15653.372	370.237 [15]		<0.0001	18.8%
varcoP	15658.552	375.417 [15]		<0.0001	19.1%
rPVIp	15498.035	214.900 [15]		<0.0001	10.9%
nPVIp	15486.248	203.113 [15]		<0.0001	10.3%

a. The χ^2 value of the final model is calculated by taking the difference between the −2log-likelihood ratios (−2LL) of the null model and the final model. The χ^2 value of each tested measure is calculated by taking the difference between the −2LLs of the final model and each reduced model.

b. The variability explained is calculated by taking the percentage of the χ^2 value of each measure over the sum of all χ^2 values for all measures ($\Sigma\chi^2$).

3.3 Speaker recognition using intensity and duration measures

Table 6 summarises the speaker-recognition results based on all eight intensity measures compared to a reduced set of measures that have been found to explain most between-speaker variability (varcoP, stdevP, varcoM, and stdevM; see Section 3.2). With a reduced set of measures, mean recognition rates dropped with increased standard errors, yet one exception occurred: the standard error of the reduced set using feedforward neural networks was 0.1 lower than the full set (see Table 6). This shows that instead of confusing the detection algorithms, the measures which contribute less to between-speaker variability also help increase speaker-recognition performance of the models. For this reason, all eight intensity measures were used in the following speaker-recognition experiments.

Table 6: Comparisons of speaker-recognition success using a full set of intensity measures and a reduced set of intensity measures as predictor variables

		Mean hit rate ± std.err (in %)	
		non-normalised score	z-score
(i) BonnTempo results			
Multinomial logistic regression	Full set[a]	50.7 ± 5.5	51.1 ± 4.6
	Reduced set[b]	46.4 ± 5.7	48.5 ± 5.8
Feedforward neural networks	Full set[a]	44.1 ± 9.0	44.1 ± 4.6
	Reduced set[b]	34.6 ± 8.9	41.7 ± 12.0
(ii) TEVOID results			
Multinomial logistic regression	Full set[a]	38.3 ± 4.6	46.1 ± 4.1
	Reduced set[b]	35.4 ± 8.9	40.6 ± 10.1
Feedforward neural networks	Full set[a]	32.2 ± 4.5	40.3 ± 4.9
	Reduced set[b]	31.0 ± 7.7	35.4 ± 8.9

a. The full set of measures includes stdevM, varcoM, rPVIm, nPVIm, stdevP, varcoP, rPVIp and nPVIp.
b. The reduced set of measures includes stdevM, varcoM, stdevP and varcoP, which explained most between-speaker variability (see Section 3.2).

Figure 5 displays average speaker-recognition rates (± standard errors) yielded from both multinomial logistic regressions and neural networks using intensity measures and/or duration measures (summarised in Table 1) with (or without) z-score normalisations in both BonnTempo and TEVOID databases. The numbers of speakers who have been correctly recognised above chance levels in both databases are reported in Table 7.

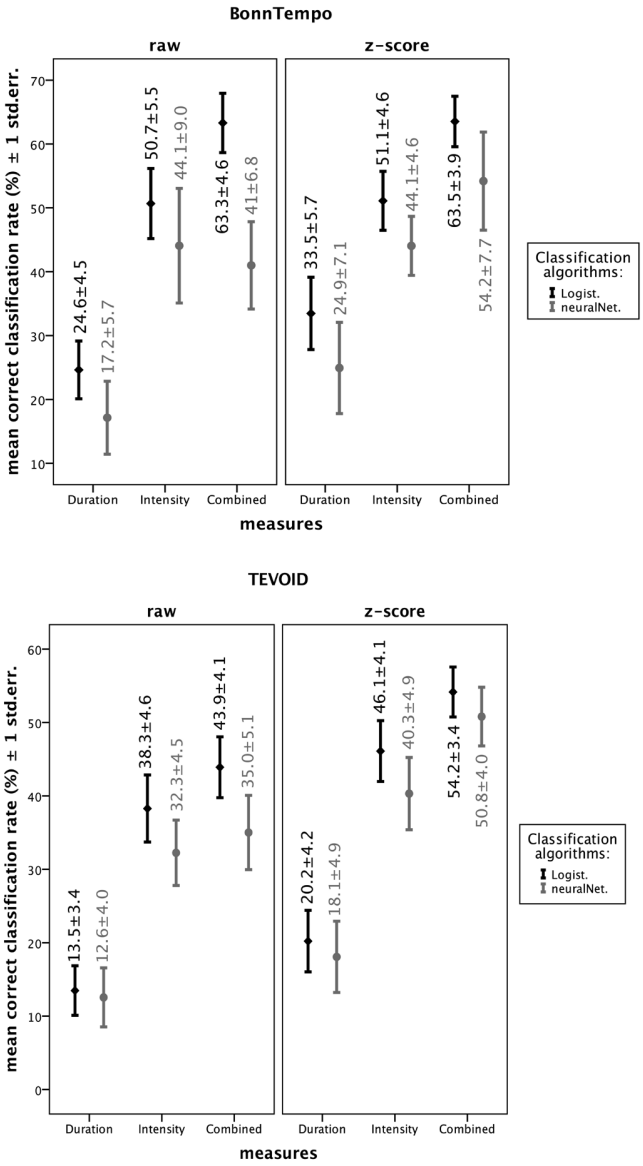


Figure 5: The mean speaker-recognition rates using both original and z-score normalised measures (duration, intensity and combined) yielded from the multinomial logistic regression (left bars) and feedforward neural networks (right bars) for the BonnTempo (top) and the TEVOID (bottom) databases.

Table 7: The numbers of speakers who have been correctly recognised above chance levels using both non-normalised and z-score normalised measures (duration, intensity and combined) yielded from the multinomial logistic regression and feedforward neural networks for the BonnTempo and the TEVOID data

	Multinomial logistic regression	Feedforward neural networks
(i) BonnTempo database[a]		
Duration (non-normalised)	10	7
Duration (z-score)	11	9
Intensity (non-normalised)	12	10
Intensity (z-score)	12	12
Combined (non-normalised)	12	12
Combined (z-score)	12	11
(ii) TEVOID database[b]		
Duration (non-normalised)	11	9
Duration (z-score)	13	11
Intensity (non-normalised)	16	16
Intensity (z-score)	16	16
Combined (non-normalised)	16	16
Combined (z-score)	16	16

a. Chance level for BonnTempo database: 100% ÷ 12 speakers ≃ 8.3%
b. Chance level for TEVOID database: 100% ÷ 16 speakers ≃ 6.3%

For the 3 (MEASURE TYPE) × 2 (RECOGNITION ALGORITHM) × 2 (NORMALISATION) factorial ANOVAs with recognition rate as dependent variable, no possible interactions were significant for both databases (all p values > 0.5) which means that main effects were readily interpretable. The main effect of MEASURE TYPE (intensity * duration * combined) was significant in both BonnTempo ($F_{(2, 132)} = 27.68$, $p < 0.0001$) and TEVOID ($F_{(2, 180)} = 53.67$, $p < 0.0001$). Post hoc comparisons with Tukey HSD adjustments showed that intensity measures and combined measures were significantly better than duration measures alone for both BonnTempo and TEVOID (adjusted $p < 0.0001$). However, the differences between intensity and combined measures for both corpora were not significant (TEVOID: adjusted $p = 0.07$; BonnTempo: adjusted $p = 0.15$). The main effect of RECOGNITION ALGORITHM (multinomial logistic regression * neural networks) in the case of BonnTempo showed that multinomial logistic regressions performed highly significantly better than feedforward neural networks ($F_{(1, 132)} = 8.70$, $p < 0.004$). For TEVOID, however, the choice of algorithms showed only a tendency ($F_{(1, 180)} = 3.34$, $p = 0.07$). The main effect of NORMALISATION (non-normalised *

z-score) showed that z-score measures performed highly significantly better in TEVOID ($F_{(1, 180)} = 13.28$, $p < 0.0001$), but the effect was not significant in BonnTempo ($F_{(1, 132)} = 2.145$, $p = 0.15$).

For the syllable duration measures (ΔSyll(ln) and nPVISyll; not plotted) we found that they yielded recognition rates of 11.2% (using neural network) and 13.6% (using multinomial logistic regression) for the BonnTempo database, and 8.6% (using neural network) and 8.8% (using multinomial logistic regression) for the TEVOID database. Combining the syllable duration measures with the ones in Dellwo, Leemann and Kolly (2015), the recognition rates were not much different than using the latter alone: 21.0% (using neural networks) and 27.9% (using multinomial logistic regression) for BonnTempo; and 12.0% (using neural networks) and 13.5% (using multinomial logistic regression) for TEVOID. This demonstrated that syllable durations do not contribute to speaker-recognition performance in our models.

4. Discussion

Our study showed that measures of intensity variability vary strongly and consistently between speakers (Section 3.2). Most importantly, however, in a semi-automatic speaker-recognition experiment the intensity measures lead to significantly higher performance in both databases (BonnTempo and TEVOID) compared to the duration measures in Dellwo, Leemann and Kolly (2015). This was true for both recognition models (logistic and neural networks) and for non-normalised or z-score normalised data. Although a higher recognition performance with the combined measures was observed (see Figure 4), this effect only showed a tendency in TEVOID ($p < 0.1$) and no effect in BonnTempo ($p > 0.15$). We take our result as support for the view that suprasegmental intensity variability contains more speaker-specific information than suprasegmental durational information. It is thus possible that perceptual rhythmic differences between speakers are more grounded in suprasegmental intensity characteristics rather than durational characteristics. Since a high correlation between the degree of mouth aperture and intensity fluctuations as a function of time has been found (e.g., Chandrasekaran et al. 2009), it is plausible that time-integrated intensity variability measures encode more information of speaker idiosyncratic articulatory behaviour than interval duration alone. Therefore, intensity-based rhythm measures may add more orthogonal dimensions to the feature space that characterise individual speakers. We envisage an improved speaker-recognition performance when such measures are implemented in (semi)automatic speaker-recognition systems or other environments where speaker-specific information is relevant (e.g., forensic speaker comparison).

Numerous duration-based rhythm measures have been developed over the past decades. The intervals on which measurements were based extended from the vocalic or consonantal interval duration (Ramus, Nespor and Mehler 2009; Grabe and Low 2002) to, for example, syllable or foot duration (e.g., Nolan and Asu 2009), successive vocalic-consonantal interval duration (Liss et al. 2009) and voiced or unvoiced interval duration (Dellwo and Fourcin 2013). We adopted the ones that have already been attested to show significant between-speaker effect (Dellwo, Leemann and Kolly 2015). Although we also tested syllable-sized duration measures, they in fact showed very limited contributions. Another set of duration measures also showing significant between-speaker effect (Leemann, Kolly and Dellwo 2014) have been compared with the intensity measures in terms of speaker-recognition strength as well (He, Glavitsch and Dellwo 2015a). Similar results were obtained.

The usefulness of z-score normalisation is more salient in a larger database. In TEVOID, sentence variability (number of sentences = 256) is strong, and z-score normalisation successfully reduced such effect, and speaker-recognition success was significantly better using z-score normalised measures ($p < 0.0001$). This implies that, if these rhythm measures are implemented in an automatic speaker verification system with a large fixed sentence bank, z-score normalisation by sentence should be helpful to increase recognition rates.

The selection of recognition algorithms also plays a role on how successful speaker recognitions are, but it depends on the database being analysed. In BonnTempo, the multinomial logistic regression performed significantly better than the neural networks ($p < 0.004$), but the algorithms did not differ significantly in TEVOID, although a tendency that the multinomial logistic regression worked better was observed ($p < 0.1$). This might be because TEVOID is much larger (16 speakers × 256 sentences) than BonnTempo (12 speakers × 7 sentences × 5 tempi) and therefore permit more adequate training. Between-algorithm recognition differences may be minimised with more adequate training data. Moreover, the result also implies that the choice of algorithms for speaker recognition is important. Optimising existing algorithms or developing new algorithms to maximise speaker-recognition success using intensity measures could thus be part of further research.

Findings on the performance of individual intensity measures are also interesting. For both databases, varcoP, stdevP, varcoM and stdevM explained the most between-speaker variability. Such measures evaluate the holistic intensity variability in the signal, instead of the local, sequential differences between syllables as measured by r(n)PVIm(p). The reason why such sequential measures did not explain more between-speaker variability may be a formal problem related to the

PVI (Gibbon 2003). Gibbon showed that for tuples like (2, 4, 2, 4, 2, 4) and (2, 4, 8, 16, 32, 64), both have identical PVI values, but differ drastically in terms of their standard deviations and variation coefficients. Although such extreme cases are rare in real speech, it is well possible that local averaging might reduce the variability under observation. This is also true for duration measures of between-speaker rhythmic differences. In Wiget et al. (2010), the PVI of vocalic interval durations did not show significant speaker effects, but the percentage of vocalic durations (%V) and variation coefficient of vocalic durations (varcoV) all showed such effects. This is further evidence for the view that the nature of PVI computations reduces measured variability between speakers. For a larger database like TEVOID with 256 sentences per speaker, the probability that individual differences are reduced that way is correspondingly higher. That could explain why the total between-speaker variability explained by the PVIs is lower in TEVOID (29.4%) than in BonnTempo (45.4%), especially rPVIm and nPVIm, each explaining only 4.1% of the variability between speakers in TEVOID. Additionally, while the PVI measures may capture the syllable-to-syllable intensity structure of language, the standard deviation and varco are more sensitive to between-speaker differences in loudness variation. A speaker who favours crescendo or diminuendo towards the end of a phrase can be more easily differentiated from the one who prefers to speak at a steady volume.[2]

In addition, the peak measures (stdevP, varcoP, rPVIp and nPVIp) conjointly explained more between-speaker variability than the mean measures (stdevM, varcoM, rPVIm and nPVIm) in the TEVOID database. It is possible that this stands in relation with the articulatory rationale (Section 1.2) according to which differences in the individual anatomy of a speaker's articulators result in differences in their movements and thus in intensity patterns over time. It is possible that peak measures can better capture such idiosyncratic articulatory behaviour. When the mandible or tongue tip reaches maximum displacement, such events could find their acoustic correlates in the signal as intensity peaks. Strong alignments between jaw/tongue tip maximum displacements and intensity peaks are shown in Birkholz, Kröger and Neuschaefer-Rube (2011), Chandrasekaran et al. (2009) and Erickson et al. (2015). It is unclear whether the mean intensity of a syllable is affected by this to the same degree. In the BonnTempo database the difference between peak and mean measures was less obvious. This again might have to do with the drastic within-speaker rate variability in this database. It is unclear what effect this variability has on syllabic intensity peaks, but it should be assumed that such effects are high between very slow and very fast speech, for example.

There are also several issues that are not addressed in this article, and they warrant further investigation. It would be important to examine how and to what

extent different forms of signal degradations such as lossy audio codecs (e.g., MP3 and Ogg Vorbis) or amplitude clipping affect between-speaker intensity variability. Also, the nonlinear dynamic range compression (e.g., compressors and limiters) that is typically applied to recordings in the entertainment sector may limit variability studied in this paper. Further, it would be interesting to study the effect of variable source locations on intensity variability between speakers.

Given that some of our measures calculate proportional differences between consecutive syllables, the aforementioned influences might not be very dramatic. In addition, it seems sensible to go beyond overall intensity variability across all frequencies as we tested in this study, and look at intensity fluctuations in different frequency sub-bands. Such analysis would be particularly important to severely clipped speech, where overall intensity variability is reduced tremendously but possibly less in particular frequency bands. The fact that rhythmic information is frequency-band specific has already been pointed out by Leong and Goswami (2014) and Leong et al. (2014) in a different context.

In addition to the 'static' intensity measures we proposed in this article, dynamic changes between intensity peaks and valleys in the signals also warrant further investigations. Such intensity dynamics may also show speaker idiosyncratic characteristics (He, Glavitsch and Dellwo 2015b) and could be possible acoustic correlates of individual biomechanical characteristics of the articulators (e.g., the velocity of intensity changes vs. the velocity of articulatory movements). It would be interesting if mathematical relationships between biomechanical measurements of articulators and intensity measurements could be established.

Finally, as mentioned in the introduction, between-language rhythmic differences have been extensively investigated from the timing perspective. Testing speech rhythm with more typologically different languages using intensity-based measures would be interesting as well. We can hence gain more insights than the current literature (Cichocki, Selouani and Perreault 2014; Low 1998; He 2012) has to offer.

5. Conclusion

This study examined speaker idiosyncratic speech rhythm from the perspective of intensity variability. Results showed that, despite high within-speaker variability, between-speaker variability was significant across both databases. Moreover, we found that intensity-based rhythm measures contained more speaker-specific information than duration-based measures. Although the choice of recognition algorithms and data normalisations had significant impact on speaker recognition in one database or the other, the superiority of intensity measures remained in both databases across all conditions.

Acoustic investigations on speaker idiosyncrasy have been extensively carried out in the frequency domain to explore individual glottal source characteristics (e.g., Gudnason and Brookes 2008; Jessen, Köster and Gfroerer 2005; Jessen 2009; Leemann, Mixdorff, O'Reilly, Kolly and Dellwo 2014) and vocal tract resonance characteristics (e.g., Duckworth, McDougall, de Jong and Shockey 2011; McDougall 2006; Morrison 2009; Zhang, van de Weijer and Cui 2006). Individual differences from the perspective of articulatory behaviour have been studied in terms of rhythm only recently (Dellwo, Leemann and Kolly 2012, 2015; Leemann, Kolly and Dellwo 2014). This study adds more evidence of speaker-specific articulation from the intensity dimension. Such intensity measures may add more orthogonal dimensions to the speaker feature space, which may facilitate tasks such as speaker recognition or forensic speaker comparison.

Acknowledgments

This study was supported by the Gebert Rüf Foundation (Grant No. GRS-027/13) and the Swiss National Science Foundation (Grant No. 100015_135287). The authors would like to thank Adrian Leemann and Marie-José Kolly for their work on the TEVOID corpus, Petra Wagner, Ingmar Steiner and Bianca Aschenberner for their work on the BonnTempo corpus, Ulrike Glavitsch for helpful comments on the paper, Sandra Schwab, Elise Dupuis Lozeron and Ben Bolker for helpful comments on statistics.

About the authors

Lei He (MA, MSc, Dr. des.) is a postdoctoral researcher in the Phonetics Lab at University of Zurich (UZH). He studied in the Doctoral Programme in Linguistics (DPL) at UZH, and successfully defended his dissertation on speaker idiosyncratic intensity variability in the speech signal (summa cum laude). He is interested in the rhythmicity of speech, in particular how articulatory factors affect the acoustic cues that underpin rhythmic differences between speakers.

Volker Dellwo (MA, PhD) is Assistant Professor of Phonetics and Speech Sciences in the Phonetics Lab at University of Zurich (UZH) and occasionally works as an expert witness in forensic phonetics. His research interests lie in a wide variety of duration, rhythm and timing phenomena in speech, typically in relation to speaker individuality. He is the principal investigator in two major grant-funded research projects addressing temporal aspects in speaker individuality.

Notes

1. Semiautomatic speaker recognition refers to a method (or a group of methods) of speaker recognition that operates partially automatically and partially manually in the central processing stages. Specifically, manual processing

occurs at the feature extraction level. However, in automatic speaker recognition all central processing stages are accomplished automatically (Drygajlo, Jessen, Gfroerer, Wagner, Vermeulen and Niemi 2015). In our study, the syllabification of both databases involved extensive human interventions; therefore, it falls within semiautomatic speaker recognition. We thank an editor for the clarification of terminology use.

2. We thank a reviewer for pointing this interpretation out, as this may shed light on the functions of the two classes (holistic and local, see Table 1(i)) of intensity measures in addition to what we introduced in Section 1.3. Holistic measures may capture the habitual crescendo or diminuendo speaking idiosyncrasy better than local measures.

References

Abercrombie, D. (1967) *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

Arvaniti, A. (2012) The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics* 40(3): 351–373. http://dx.doi.org/10.1016/j.wocn.2012.02.003

Baayen, R. H. (2008) *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511801686

Barbosa, P. A. (2002) Explaining cross-linguistic rhythmic variability via a coupled-oscillator model of rhythm production. In *Proceedings of Speech Prosody 2002* 163–166. Aix-en-Provence, France.

Bates, D., Maechler, M., Bolker, B. and Walker, S. (2014) *lme4: Linear mixed-effects models using Eigen and S4* (R package version 1.1-7). http://CRAN.R-project.org/package=lme4 Accessed 19 October 2014.

Birkholz, P., Kröger, B. J. and Neuschaefer-Rube, C. (2011) Model-based reproduction of articulatory trajectories for consonant-vowel sequences. *IEEE Transactions on Audio, Speech, and Language Processing* 19(5): 1422–1433. http://dx.doi.org/10.1109/TASL.2010.2091632

Boersma, P. (1998) *Vector value interpolation* (Praat online manual). http://www.fon.hum.uva.nl/praat/manual/vector_value_interpolation.html Accessed 11 October 2015.

Boersma, P. and Weenink, D. (2014) *Praat: Doing Phonetics by Computer* (version 5.3.65). http://www.praat.org Accessed 2 March 2014.

Bolker, B. J. (2015) Linear and generalized linear mixed models. In G. A. Fox, S. Negrete-Yankelevich and V. J. Sosa (eds) *Ecological Statistics: Contemporary Theory and Application* 309–333. Oxford: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199672547.003.0014

Byrne, C. and Foulkes, P. (2004) The 'mobile phone effect' on vowel formants. *International Journal of Speech, Language and the Law* 11(1): 83–102. http://dx.doi.org/10.1558/ijsll.v11i1.83

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A. and Ghazanfar, A. A. (2009) The natural statistics of audiovisual speech. *PLoS Computational Biology* 5: e1000436. http://dx.doi.org/10.1371/journal.pcbi.1000436

Cichocki, W., Selouani, S.-A. and Perreault, Y. (2014) Measuring rhythm in dialects of New Brunswick French: Is there a role for intensity? *Canadian Acoustics* 43(3): 90–91.

Cumming, R. (2011) Perceptually-informed quantification of speech rhythm in pairwise variability indices. *Phonetica* 68(4): 256–277. http://dx.doi.org/10.1159/000335416

Cummins, F. and Port, R. (1998) Rhythmic constraints on stress timing in English. *Journal of Phonetics* 26(2): 145–171. http://dx.doi.org/10.1006/jpho.1998.0070

Dellwo, V. (2006) Rhythm and speech rate: a variation coefficient for deltaC. In P. Karnowski and I. Szigeti (eds) *Language and Language Processing* 231–241. Frankfurt am Main: Peter Lang. http://dx.doi.org/10.5167/uzh-111789

Dellwo, V. (2009) Choosing the right rate normalization method for measurements of speech rhythm. In S. Schmid, M. Schwarzenbach and D. Studer (eds) *La dimensione temporale del parlato: Atti del 5° Convegno Nazionale AISV 2009* 13–32. Torriana: EDK Editore.

Dellwo, V. (2010) Influences of speech rate on the acoustic correlates of speech rhythm: an experimental phonetic study based on acoustic and perceptual evidence. Unpublished PhD dissertation, Bonn University.

Dellwo, V. (2015) What does voice and silence tell us about speaker identity? An introduction to temporal speaker individualities and their use for forensic speaker comparison. In G. M. Schnider, M. C. Janner and B. Élie (eds) *Vox & Silentium* 17–35. Bern: Peter Lang. http://dx.doi.org/10.3726/978-3-0351-0823-1

Dellwo, V. and Fourcin, A. (2013) Rhythmic characteristics of voice between and within languages. *Travaux Neuchâtelois de Linguistique* 59: 87–107. http://dx.doi.org/10.5167/uzh-91230

Dellwo, V., Leemann, A. and Kolly, M.-J. (2012) Speaker idiosyncratic rhythmic features in the speech signal. In *Proceedings of INTERSPEECH 2012* 1582–1585. Portland, USA. http://dx.doi.org/10.5167/uzh-68554

Dellwo, V., Leemann, A. and Kolly, M.-J. (2015) Rhythmic variability between speakers: articulatory, prosodic, and linguistic factors. *Journal of the Acoustical Society of America* 137: 1513–1528. http://dx.doi.org/10.1121/1.4906837

Dellwo, V., Steiner, I., Aschenberner, B., Dankovičova, J. and Wagner, P. (2004) The BonnTempo-Corpus & BonnTempo-Tools: a database for the study of speech rhythm and rate. In *Proceedings of the 8th ICSLP/INTERSPEECH 2004* 777–780. Jeju Island, Korea.

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J. and Niemi, T. (2015) *Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition*. Frankfurt am Main: Verlag für Polizeiwissenschaft.

Duckworth, M., McDougall, K., de Jong, G. and Shockey, L. (2011) Improving the consistency of formant measurement. *International Journal of Speech, Language and the Law* 18(1): 35–51. http://dx.doi.org/10.1558/ijsll.v18i1.35

Erickson, D., Kim, J., Kawahara, S., Wilson, I., Menezes, C., Suemitsu, A. and Moore, J. (2015) Bridging articulation and perception: the C/D model and contrastive emphasis. In *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)* 0527.1–5. Glasgow, UK.

Eriksson, A. and Wretling, P. (1997) How flexible is the human voice? – A case study of mimicry. In *Proceedings of EUROSPEECH 1997* 1043–1046. Rhodes, Greece.

Frota, S. and Vigário, M. (2001) On the correlates of rhythmic distinctions: the European/Brazilian Portuguese case. *Probus* 13(2): 247–275. http://dx.doi.org/10.1515/prbs.2001.005

Fuchs, R. (2014) Integrating variability in loudness and duration in a multidimensional model of speech rhythm: evidence from Indian English and British English. In *Proceedings of Speech Prosody 2014* 290–294. Dublin, Ireland.

Galves, A., Garcia, J., Duarte, D. and Galves, C. (2002) Sonority as a basis for rhythmic class discrimination. In *Proceedings of Speech Prosody 2002* 323–326. Aix-en-Provence, France.

Garnier, M., Wolfe, J., Henrich, N. and Smith, J. (2008) Interrelationship between vocal effort and vocal tract acoustics: a pilot study. In *Proceedings of INTERSPEECH 2008* 2302–2305. Brisbane, Australia.

Gibbon, D. (2003) Computational modelling of rhythm as alternation, iteration and hierarchy. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)* 2489–2492. Barcelona, Spain.

Grabe, E. and Low E. L. (2002) Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven and N. Warner (eds) *Laboratory Phonology 7* 514–546. Berlin: Mouton de Gruyter. http://dx.doi.org/10.1515/9783110197105.515

Gudnason, J. and Brookes, M. (2008) Voice source cepstrum coefficients for speaker identification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008)* 4821–4824. Las Vegas, USA. http://dx.doi.org/10.1109/ICASSP.2008.4518736

He, L. (2012) Syllabic intensity variations as quantification of speech rhythm: evidence from both L1 and L2. In *Proceedings of Speech Prosody 2012* 466–469. Shanghai, China.

He, L. and Dellwo, V. (2014) Speaker idiosyncratic variability of intensity across syllables. In *Proceedings of INTERSPEECH 2014* 233–237. Singapore. http://dx.doi.org/10.5167/uzh-103024

He, L., Glavitsch, U. and Dellwo, V. (2015a) Comparisons of speaker recognition strengths using suprasegmental duration and intensity variability: an artificial neural networks approach. In *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)* 0395.1–5. Glasgow, UK.

He, L., Glavitsch, U. and Dellwo, V. (2015b) Inter-speaker variability in intensity dynamics. Paper presented at the 24th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA2015), 8–10 July 2015, Leiden, the Netherlands.

IBM Corp. (2013) IBM SPSS Statistics for Macintosh (version 22.0). Armonk, NY: International Business Machines Corporation.

Jessen, M. (2009) Forensic phonetics and the influence of speaking style on global measures of fundamental frequency. In G. Grewendorf and M. Rathert (eds) *Formal Linguistics and Law* 115–139. Bern: Mouton de Gruyter. http://dx.doi.org/10.1515/9783110218398.2.115

Jessen, M., Köster, O. and Gfroerer, S. (2005) Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech, Language and the Law* 12(2): 174–213. http://dx.doi.org/10.1558/sll.2005.12.2.174

Jovičić, S., Jovanović, N., Subotić, M. and Grozdić, Đ. (2015) Impact of mobile phone usage on speech spectral features: some preliminary findings. *International Journal of Speech, Language and the Law* 22(1): 111–125. http://dx.doi.org/10.1558/ijsll.v22i1.17880

Kohler, K. J. (2008) The perception of prominence patterns. *Phonetica* 65(4): 257–269. http://dx.doi.org/10.1159/000192795

Kohler, K. J. (2009) Rhythm in speech and language: a new research paradigm. *Phonetica* 66(1–2): 29–45. http://dx.doi.org/10.1159/000208929

Kolly, M.-J. and Dellwo, V. (2014) Cues to linguistic origin: the contribution of speech temporal information to foreign accent recognition. *Journal of Phonetics* 42: 12–23. http://dx.doi.org/10.1016/j.wocn.2013.11.004

Krehbiel, T. C. (2004) Correlation coefficient rule of thumb. *Decision Sciences* 2(1): 97–100. http://dx.doi.org/10.1111/j.0011-7315.2004.00025.x

Lee, C. S., Kitamura, C., Burnham, D. and Todd, N. P. M. (2014) On the rhythm of infant-versus adult-directed speech in Australian English. *Journal of the Acoustical Society of America* 136: 357–365. http://dx.doi.org/10.1121/1.4883479

Lee, C. S. and Todd, N. P. M. (2004) Towards an auditory account of speech rhythm: application of a model of the auditory 'primal sketch' to two multi-language corpora. *Cognition* 93(3): 225–254. http://dx.doi.org/10.1016/j.cognition.2003.10.012

Leemann, A. and Kolly, M.-J. (2015) Speaker-invariant suprasegmental temporal features in normal and disguised speech. *Speech Communication* 75: 97–122. http://dx.doi.org/10.1016/j.specom.2015.10.002

Leemann, A., Kolly, M.-J. and Dellwo, V. (2014) Speech-individuality in suprasegmental temporal features: implications for forensic voice comparison. *Forensic Science International* 238: 59–67. http://dx.doi.org/10.1016/j.forsciint.2014.02.019

Leemann, A., Mixdorff, H., O'Reilly, M., Kolly, M.-J. and Dellwo, V. (2014) Speaker-individuality in Fujisaki model f0 features: implications for forensic voice comparison. *International Journal of Speech, Language and the Law* 21(2): 343–370. http://dx.doi.org/10.1558/ijsll.v21i2.343

Lenth, R. V. (2014) *lsmeans: Least-Squares Means* (R package version 2.12). http://CRAN.R-project.org/package=lsmeans/ Accessed 21 October 2014.

Leong, V. and Goswami, U. (2014) Impaired extraction of speech rhythm from temporal modulation patterns in speech in developmental dyslexia. *Frontiers in Human Neuroscience* 8: 96. http://dx.doi.org/10.3389/fnhum.2014.00096

Leong, V., Stone, M. A., Turner, R. E. and Goswami, U. (2014) A role for amplitude modulation phase relationships in speech rhythm perception. *Journal of the Acoustical Society of America* 136: 366–381. http://dx.doi.org/10.1121/1.4883366

Licklider, J. C. R. and Pollack, I. (1948) Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech. *Journal of the Acoustical Society of America* 20: 42–51. http://dx.doi.org/10.1121/1.1906346

Liss, J. M., White, L., Mattys, S., Lansford, K., Lotto, A. J., Spitzer, S. M. and Caviness, J. N. (2009) Quantifying speech rhythm abnormalities in the dysarthrias. *Journal of Speech Language and Hearing Research* 52: 1334–1352. http://dx.doi.org/10.1044/1092-4388(2009/08-0208)

Loukina, A., Kochanski, G., Rosner, B., Keane, E. and Shih, C. (2011) Rhythm measures and dimensions of durational variation in speech. *Journal of the Acoustical Society of America* 129: 3258–3270. http://dx.doi.org/10.1121/1.3559709

Low, E. L. (1998) Prosodic prominence in Singapore English. Unpublished PhD thesis, University of Cambridge.

Low, E. L., Grabe, E. and Nolan, F. (2000) Quantitative characterization of speech rhythm: syllable-timing in Singapore English. *Language and Speech* 43(4): 377–401. http://dx.doi.org/10.1177/00238309000430040301

McDougall, K. (2006) Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies. *International Journal of Speech, Language and the Law* 13(1): 89-126. http://dx.doi.org/10.1558/ijsll.v13i1.89

Morrison, G. (2009) Likelihood-ratio-based forensic speaker comparison using parametric representations of the formant trajectories of diphthongs. *Journal of the Acoustical Society of America* 125: 2387–2397. http://dx.doi.org/10.1121/1.3081384

Nolan, F. and Asu, E. L. (2009) The pairwise variability index and coexisting rhythms in language. *Phonetica* 66(1–2): 64–77. http://dx.doi.org/10.1159/000208931

Nolan, F. and Jeon, H.-S. (2015) Speech rhythm: a metaphor? *Philosophical Transactions of the Royal Society B* 369: 20130396. http://dx.doi.org/10.1098/rstb.2013.0396

O'Dell, M. L. and Nieminen, T. (1999) Coupled oscillator model of speech rhythm. In *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS)* 1075–1078. San Francisco, USA.

Payne, E., Post, B., Astruc, L., Prieto, P. and Vanrell, M. (2012) Measuring child rhythm. *Language and Speech* 55(2): 203–229. http://dx.doi.org/10.1177/0023830911417687

Perrier, P. (2012) Gesture planning integrating knowledge of the motor plant's dynamics: a literature review from motor control and speech motor control. In S. Fuchs, M. Weirich, D. Pape and P. Perrier (eds) *Speech Planning and Dynamics* 191–238. Frankfurt am Main: Peter Lang. http://dx.doi.org/10.3726/978-3-653-01438-9

Plant, R. L. and Younger, R. M. (2000) The interrelationship of subglottic air pressure, fundamental frequency, and vocal intensity during speech. *Journal of Voice* 14(2): 170–177. http://dx.doi.org/10.1016/S0892-1997(00)80024-7

Polyanskaya, L. and Ordin, M. (2015) Acquisition of speech rhythm in first language. *Journal of the Acoustical Society of America* 138: 199–204. http://dx.doi.org/10.1121/1.4929616

R Core Team (2014) *R: A Language and Environment for Statistical Computing* (version 3.1.1). R Foundation for Statistical Computing, Vienna. http://www.R-project.org Accessed 22 July 2014.

Ramus, F., Nespor, M. and Mehler J. (1999) Correlates of linguistic rhythm in the speech signal. *Cognition* 73(3): 265–292. http://dx.doi.org/10.1016/S0010-0277(00)00101-3

Rathcke, T. V. and Smith, R. H. (2015) Speech timing and linguistic rhythm: on the acoustic bases of rhythm typologies. *Journal of the Acoustical Society of America* 137: 2834–2845. http://dx.doi.org/10.1121/1.4919322

Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A. and Stolcke, A. (2005) Modeling prosodic feature sequences for speaker recognition. *Speech Communication* 46(3–4): 455–472. http://dx.doi.org/10.1016/j.specom.2005.02.018

Tilsen, S. and Arvaniti, A. (2013) Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages. *Journal of the Acoustical Society of America* 134: 628–639. http://dx.doi.org/10.1121/1.4807565

Tilsen, S. and Johnson, K. (2008) Low-frequency Fourier analysis of speech rhythm. *Journal of the Acoustical Society of America* 124: EL34–EL39. http://dx.doi.org/10.1121/1.2947626

White, L., Liss, J. and Dellwo, V. (2010) Assessment of rhythm. In A. Lowit and R. D. Kent (eds) *Assessment of Motor Speech Disorders* 312–352. San Diego: Plural Publishing.

White, L., Payne, E. and Mattys, S. (2009) Rhythmic and prosodic contrast in Venetan and Sicilian Italian. In M. Vigário, S. Frota and M. J. Freitas (eds) *Phonetics and Phonology: Interactions and Interrelations* 137–158. Amsterdam: John Benjamins. http://dx.doi.org/10.1075/cilt.306.07whi

Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O. and Mattys, S. (2010) How stable are acoustic metrics of contrastive speech rhythm? *Journal of the Acoustical Society of America* 127: 1559–1569. http://dx.doi.org/10.1121/1.3293004

Wilson, J. V. and Leeper, H. A. (1992) Changes in laryngeal airway resistance in young adult men and women as a function of vocal sound pressure level and syllable context. *Journal of Voice* 6(3): 235–245. http://dx.doi.org/10.1016/S0892-1997(05)80148-1

Wretling, P. and Eriksson, A. (1998) Is articulatory timing speaker specific? – evidence from imitated voices. In P. Branderund and H. Traunmüller (eds) *Proceedings of FONETIK '98* 48–51. Stockholm: Stockholm University.

Yoon, T. J. (2010) Capturing inter-speaker invariance using statistical measures of speech rhythm. In *Proceedings of Speech Prosody 2010* 100201.1–4. Chicago, USA.

Zhang, C., van de Weijer, J. and Cui, J. (2006) Intra- and inter-speaker variations of formant pattern for lateral syllables in Standard Chinese. *Forensic Science International* 158(2–3): 117–124. http://dx.doi.org/10.1016/j.forsciint.2005.04.043

Appendix A: Reading materials in the corpora

1. BonnTempo

The vertical strokes demarcate syntactic chunks for which the intensity measures were calculated. The chunks are referred to as 'sentences' in the article.

Am nächsten Tag fuhr ich nach Husum. | Es ist eine Fahrt ans Ende der Welt; | hinter Gießen werden die Berge und Wälder eintönig, | hinter Kassel die Städte ärmlich, | und bei Salzgitter wird das Land flach und öde. | Wenn bei uns Dissidenten verbannt würden, | würden sie ans Steinhuder Meer verbannt.

2. TEVOID

The first 15 of the 256 sentences are listed in non-standard orthography aiming to capture the pronunciation of Zürich German:

1. So s Typische was sich d Lüüt vorschteled isch Kurator.
2. Ich han Freiziit.
3. Ich han käi äigeni Band.
4. Ich bin wäge Spraachwüsseschaft dänn usegheit.
5. Das han i cool gfunde.
6. Mini Mueter isch ä no nie z Wien gsi.
7. Dänn mues ich ä no überlegge, was mis nöie Hauptfach wird.
8. Ich ha jetz äifach vergliichendi Spraachwüsseschafte gno.
9. Ich ha mich ä nie würklich beworbe.
10. Wänn ich halt im Usland wär, hett ich das zmindescht mal für es Semeschter nöd.
11. Chasch ja nöd nöime andersch go studiere mit Erasmus.
12. Si liit det am Bode.
13. Zwar isch das Ganze im ne fiktive Königriich.
14. Ich wäis noöd werum si so abglänkt isch.
15. Säge mer emaal ich fahr uf Oerlike.

Appendix B: R codes used in this article

```
#01   bt stands for BonnTempo dataset; tv stands for TEVOID dataset
02    MDL1 = lmer(nPVIm ~ tempo + (1|speaker) + (1|sentence), data = bt, REML = F)
03    MDL2 = lmer(nPVIm ~ tempo + (1|sentence), data = bt, REML = F)
04    MDL3 = lmer(nPVIm ~ 1 + (1|speaker) + (1|sentence), data = bt, REML = F)
05    MDL11 = lmer(nPVIp ~ tempo + (1|speaker) + (1|sentence), data = bt, REML = F)
06    MDL22 = lmer(nPVIp ~ tempo + (1|sentence), data = bt, REML = F)
07    MDL33 = lmer(nPVIp ~ 1 + (1|speaker) + (1|sentence), data = bt, REML = F)
08    MDL4 = lmer(nPVIm ~ 1 + (1|speaker) + (1|sentence), data = tv, REML = F)
09    MDL5 = lmer(nPVIm ~ 1 + (1|sentence), data = tv, REML = F)
10    MDL44 = lmer(nPVIp ~ 1 + (1|speaker) + (1|sentence), data = tv, REML = F)
11    MDL55 = lmer(nPVIp ~ 1 + (1|sentence), data = tv, REML = F)
12    anova(MDL1, MDL2)
13    anova(MDL1, MDL3)
14    anova(MDL11, MDL22)
15    anova(MDL11, MDL33)
16    anova(MDL4, MDL5)
17    anova(MDL44, MDL55)
18    lsmeans(MDL1, pairwise ~ tempo, adjust = 'tukey')
```

Lines 02–11 fit models in accordance with Table 3. Lines 12–17 perform the likelihood ratio tests; the results are shown in Table 4. Line 18 compares the least-squares means between tempo versions for nPVIm in BonnTempo.

Study IV

Between-speaker variability in temporal organizations of intensity contours

This work has been published in May 2017 in the *Journal of the Acoustical Society of America* (Volume 141, Issue 5, pp. EL488-EL494).
Official URL (DOI):
http://dx.doi.org/10.1121/1.4983398

Lei He and Volker Dellwo: JASA Express Letters [http://dx.doi.org/10.1121/1.4983398] Published Online 18 May 2017

Between-speaker variability in temporal organizations of intensity contours

Lei He[a] **and Volker Dellwo**
Phonetics Laboratory, Institute of Computational Linguistics, University of Zurich, Andreasstrasse 15, CH-8050 Zurich, Switzerland
lei.he@uzh.ch, volker.dellwo@uzh.ch

Abstract: Intensity contours of speech signals were sub-divided into positive and negative dynamics. Positive dynamics were defined as the speed of increases in intensity from amplitude troughs to subsequent peaks, and negative dynamics as the speed of decreases in intensity from peaks to troughs. Mean, standard deviation, and sequential variability were measured for both dynamics in each sentence. Analyses showed that measures of both dynamics were separately classified and between-speaker variability was largely explained by measures of negative dynamics. This suggests that parts of the signal where intensity decreases from syllable peaks are more speaker-specific. Idiosyncratic articulation may explain such results.

[AL]
Date Received: December 12, 2016 **Date Accepted:** April 28, 2017

1. Introduction

Source signals, vocal tract resonances, and articulatory movements are the essential processes of speech production. Each of these processes encodes speaker-specific information.[1] This study investigated how between-speaker differences are reflected in temporal organizations of intensity contours in terms of intensity dynamics. Intensity dynamics were defined as the speed of increase in intensity from an amplitude envelope trough point to a consecutive peak point (henceforth, positive dynamics) and the speed of decrease in intensity from a peak to a consecutive trough point (henceforth, negative dynamics).

Speaker idiosyncratic characteristics in both glottal vibrations and vocal tract resonances have been extensively studied in forensic phonetics and automatic speaker recognition (see Eriksson[3] and Kinnunen and Li[4] for reviews). Far less attention has been paid to the temporal characteristics of speech that are a result of the movements of the articulators over time.[5–7] The rationale of these studies is that articulatory movements are comparable to other domains of human movements (e.g., gait and typing) where individual differences are conspicuous.[3,5–7] Such individualities are related to both individual neurological dispositions, which constrain the motor control over the respective body parts,[8] and ontogenetic anatomical characteristics of moving body parts, which shape their biomechanical properties.[9] As a specialized domain of human motor behavior, articulation also reflects speaker individualities because of anatomical idiosyncrasies of the articulators and the way speakers acquired control over them. These result in speaker-specific articulatory kinematics, including velocity, acceleration and spatial displacement.[10,11] Such kinematic characteristics are assumed to be the reason for speaker-specific production of prosodic duration[5,6] and intensity variabilities.[7,12] The present research underlies the assumption that the intensity contour shape might be closely related to the articulatory movements responsible for the changes of mouth opening area in an utterance. Such a view is supported by Summerfield[13] who held that the amplitude envelope co-varied with the area of mouth opening, and Chandrasekaran *et al.*[2] who reported strong empirical evidence for Summerfield's claim. This suggests that intensity dynamics are strongly associated with articulatory movements that have direct influence on the speed by which the mouth opening area increases and decreases. Provided that this relationship exists and given the fact that articulatory movements vary between speakers, we hypothesize that intensity dynamics should also vary between speakers. This hypothesis was tested in the present experiment.

Why should we separate the intensity contour into positive and negative intensity dynamics? Birkholz *et al.*[14] examined the coordination between articulators by

[a]Author to whom correspondence should be addressed.

modelling both opening and closing gestures using dynamic systems. Opening and closing gestures are the articulatory movements to and from an articulatory target (typically a major turning point of articulators within a syllable). They found that the forces and motor programs acting on them in opening and closing gestures differed by their time constants. According to Ghez and Krakauer's[15] view of the motor program, the extent of a movement is planned before the movement is initiated. Speakers are therefore likely to have different motor planning for opening and closing gestures. Since the two gestures have different motor programs, it is unclear what effects this would have on the variability between speakers. For this reason, we looked at speaker-specific effects in positive and negative dynamics separately.

A series of measures was developed to capture how positive and negative intensity dynamics were distributed within utterances (Sec. 2.3). With them, we first tested whether measures of both dynamics formed into independent categories. Then, we tested whether and to what extent measures of both dynamics varied between speakers.

Why do we want to better understand speaker idiosyncratic temporal properties of the intensity contour? On the one hand there is a large theoretical interest. While indexical information has been deemed a by-product in classic linguistic theory, it is now evident that it plays a crucial role for the processing of meaning in speech communication.[16] The processes by which listeners recognize or distinguish different voices, however, are still poorly understood. Intensity contours might be factors contributing to auditory speaker recognition that have so far received hardly any attention. On the other hand, there are a variety of applications in which indexical information is of importance. In forensic voice analysis, for example, speaker comparison tasks often cannot be performed because the complexity of the acoustic correlates of voice identity within and between speakers is not yet well understood. It is thus essential to increase our knowledge beyond the classic factors like fundamental and formant frequencies or voice qualities to other acoustic domains that carry speaker-specific variation.

2. Method

2.1 Corpus

The TEVOID corpus[5,6] was used for the present study. It contains 16 native speakers of Zürich German (8 female, 8 male; mean age = 27, age standard deviation = 3.6, age range = 20–33, no reported speech and hearing disorders). They were recorded reading the same set of 256 sentences (see Fig. 1 for the distribution of sentence lengths in terms of syllable numbers) in a sound-attenuated booth (Neumann STH–100 transducer microphone (Georg Neumann GmbH, Berlin, Germany); 44.1k samples/s, 16-bit). All speakers practiced the sentences in advance to be able to read them fluently. The speakers read the sentences in a way they considered "everyday reading." Mm. 1 and Mm. 2 contain the sound files of the same sentence read by a female and a male speaker. Syllable boundaries were annotated automatically based on segment sonority rules; sonority scales were manually attributed to each segment type.[5,6]

Mm. 1. A female speaker reading the Zürich German sentence "Ich bin wäge Spraachwüsenschaft dänn usegheit." This is a file of type "wav" (266 Kb).

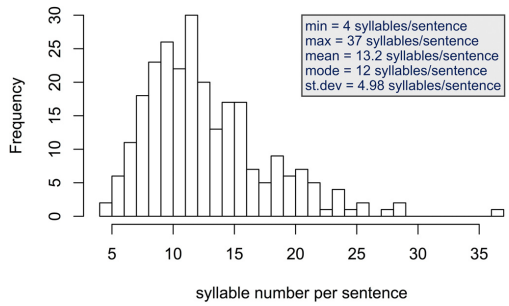


Fig. 1. (Color online) Histogram showing the distribution of sentence lengths (number of syllables per sentence) of the sentences in the TEVOID corpus.

Mm. 2. A male speaker reading the Zürich German sentence "Ich bin wäge Spraachwüsenschaft dänn usegheit." This is a file of type "wav" (291 Kb).

2.2 Extraction of the intensity contour and its peaks and troughs

We extracted the intensity contours to calculate the intensity dynamics measures (Sec. 2.3). First, the DC bias of each signal was removed by subtracting the mean amplitude. Then, the amplitude of each signal was linearly rescaled such that the maximum amplitude equated to 0.99. To obtain the intensity contour, the amplitude values of the rescaled signal were squared. A Gaussian window (approximated using the Kaiser-Bessel window: $\beta = 20$, sidelobe attenuation $\cong -190$ dB) with a length of 32 ms was multiplied repeatedly with the squared signal (window forward $= \frac{1}{4} \times 32$ ms $= 8$ ms; between-window overlap $= 75\%$). For each windowed frame, the sum of squares (SS) of the sample values was computed and substituted in $10\log_{10}\{[SS / (2 \times 10^{-5})]^2/0.032\}$ to obtain the intensity level (unit: dB re 20 μPa) in each particular frame.

Since the intensity curve obtained this way was a lower sampled function, we calculated the peak and trough points from the higher sampled amplitude envelope (obtained by low-pass filtering the full-wave rectified signal at 10 Hz [Hann filter, roll-off = 6 dB/octave]). Peak points (t_P in Fig. 2) were positioned where the envelope reached maximum values between syllable boundaries. Trough points (t_T in Fig. 2) were placed where the envelope reached minimum values between adjacent peak points. The intensity values at each peak and trough points (I_P and I_T in Fig. 2) were obtained from the intensity curve at each t_P and t_T using the cubic interpolation.

2.3 Measurement of intensity dynamics

Peak and trough points (t_P and t_T) and their associated intensity values (I_P and I_T) were obtained from each utterance. Positive dynamics ($v_I[+]$) were defined as $v_I[+] \stackrel{\text{def}}{=} (I_P - I_T)/(t_P - t_T)$, where I_P and I_T refer to the intensity values at peak and trough points represented by t_P and t_T. Similarly, negative dynamics ($v_I[-]$) were defined as $v_I[-] \stackrel{\text{def}}{=} |I_T - I_P|/(t_T - t_P)$. Absolute values were taken because we were only interested in the magnitude. Thus, we measured the speed of intensity increases and decreases. Geometrically, $v_I[+]$ and $v_I[-]$ can be demonstrated as the secant lines $\overrightarrow{I_T I_P}$ and $\overrightarrow{I_P I_T}$ in Fig. 2, and we measured the steepness of these lines.

To capture the distributions of both types of dynamics in an utterance, mean, standard deviation, and Pairwise Variability Index (PVI; for a tuple Q with n elements $\{q_1, q_2, \cdots, q_n\}$, the PVI of $Q = \sum_{i=1}^{n-1} |q_i - q_{i+1}|/(n-1)$) of both positive and negative dynamics were calculated. The PVI calculates the averaged differences between consecutive acoustic magnitudes in a speech signal (e.g. temporal intervals or here intensity dynamics).[17] It was demonstrated to be particularly suitable for summarizing the sequential variability in speech over the course of an entire utterance.[5–7,17] We notated these measures as MEAN_ $v_I[+]$, STDEV_ $v_I[+]$ and PVI_ $v_I[+]$ for positive dynamics, and MEAN_ $v_I[-]$, STDEV_ $v_I[-]$ and PVI_ $v_I[-]$ for negative dynamics. They represented

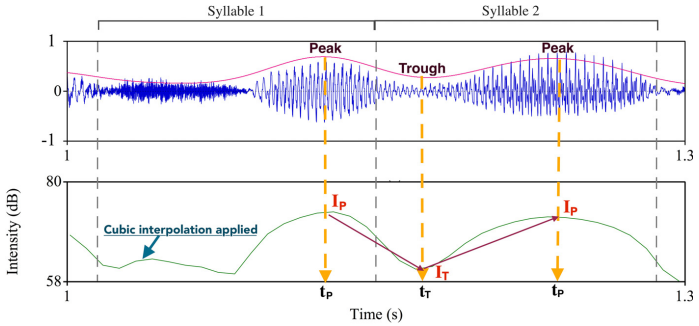


Fig. 2. (Color online) An illustration of calculating positive and negative intensity dynamics from a speech signal. The intensity contour (lower plot) was calculated from the speech waveform (upper plot). The amplitude envelope (superimposed over the waveform in the upper plot) was used to facilitate locating the peak and trough points (t_P and t_T). The peak and trough intensity values (I_P and I_T) were obtained from the intensity contour at t_P and t_T using the cubic interpolation. Intensity dynamics were calculated as how fast the intensity level dropped from a peak to its adjacent trough ($\overrightarrow{I_P I_T}$ in the lower plot, i.e., negative dynamics), or increased from a trough to its adjacent peak ($\overrightarrow{I_T I_P}$ in the lower plot, i.e., positive dynamics).

different aspects of dynamic distributions: i.e., the central tendency, the overall dispersion and sequential variability.

2.4 Statistical analyses

To control for the effect of between-sentence differences, z-score normalizations by sentence were performed for all measures of intensity dynamics: for a particular measure, the z-score of a particular sentence k was calculated as $z_k = (y_k - \bar{y}_k)/\sigma_k$, where y_k = the raw score of sentence k, $\bar{y}_k$ = the mean, and σ_k = the standard deviation of all y_k.

To test whether measures of positive and negative dynamics formed into independent categories, we performed a factor analysis (extraction method = principal components, eigenvalues ≧ 1, rotation method = Varimax with Kaiser normalization) on all measures of intensity dynamics. If measures in the two types of dynamics were classified as separate factors, we concluded that they were orthogonal and therefore encode different information.

To test the significance of between-speaker effect on each measure of intensity dynamics and the amount of between-speaker variation explained by measures of both dynamics, we employed a multinomial logistic regression (MLR). Measures of intensity dynamics were modeled as the numeric predictor variables, and speaker was modeled as the nominal response variable. Between-speaker variability explained by each measure was calculated as $(\chi^2/\Sigma\chi^2)\times 100\%$, where χ^2 refers to the likelihood ratio χ^2 of a particular measure, and $\Sigma\chi^2$ refers to the sum of likelihood ratio χ^2s of all measures.

3. Results

3.1 Factor analysis

The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy ($KMO = 0.669 > 0.5$) and the Bartlett's sphericity test ($\chi^2_{[15]} = 13249.911$, $p < 0.0005$) indicated that our dataset was suitable for factor analysis. Table 1 shows that two factors were extracted: factor 1 included all measures of negative dynamics and factor 2 included all measures of positive dynamics, suggesting that measures of both dynamics types were orthogonal.

3.2 Multinomial logistic regression

Table 2 shows the results of the MLR, examining the significance of between-speaker effect on each measure of intensity dynamics and how much between-speaker variability was explained by each measure. The negative measures collectively explained 70.35%, and the positive measures collectively explained 29.65% of between-speaker variability [Fig. 3(a)]. Figure 3(b) compares the difference between dynamics within each type of measure in explaining between-speaker variability.

4. Discussion

This paper investigated macroscopic intensity dynamics in the speech signal. Results from the MLR largely conformed to the hypothesis that intensity dynamics vary between speakers by showing that the between-speaker effect was significant in almost all measures of intensity dynamics, except PVI_ v_I[+] (see Table 2). Additionally, the amount of between-speaker variability explained by measures of both dynamics was not balanced: around 70% of between-speaker variation was explained by measures of negative dynamics. What could such a result tell us? Positive and negative dynamics

Table 1. Factor loadings matrix after Varimax rotation. The shaded loading values indicate that they are greater than the threshold (0.40), hence their associated intensity dynamics measures are classified into a particular factor.

	Factor loadings[a]	
	Factor 1	Factor 2
MEAN_ v_I[–]	0.825	0.043
STDEV_ v_I[–]	0.929	0.025
PVI_ v_I[–]	0.904	0.033
MEAN_ v_I[+]	0.098	0.780
STDEV_ v_I[+]	–0.008	0.926
PVI_ v_I[+]	0.003	0.908
Eigenvalue	2.497	2.169
% of variance explained	41.613	36.157

[a]The absolute value of a loading smaller than 0.40 indicates that the particular measure has an ignorable contribution to explaining the variance of a particular factor, and should therefore not be classified into this factor.

Table 2. Results of multinomial logistic regression.

	–2LL	$\chi^2_{[df]}$[a]	p	Variability explained[b]
(i) Model fitting information				
Null model	22713.047			
Full model	19958.848	2754.199 [90]	<0.0005	
(ii) Likelihood ratio test of each measure of intensity dynamics				
MEAN_ v_I[–]	20907.008	948.161 [15]	<0.0005	59.38%
STDEV_ v_I[–]	20100.527	141.679 [15]	<0.0005	8.88%
PVI_ v_I[–]	19992.198	33.351 [15]	<0.004	2.09%
MEAN_ v_I[+]	20304.375	345.527 [15]	<0.0005	21.64%
STDEV_ v_I[+]	20064.253	105.406 [15]	<0.0005	6.60%
PVI_ v_I[+]	19981.516	22.668 [15]	= 0.09	1.42%
		$\Sigma\chi^2$ =1596.792		Σ% = 100%

[a]The χ^2 value of the final model was calculated by taking the difference between the −2log-likelihood ratios (–2LL) of the null model and the final model. The χ^2 value of each tested measure was calculated by taking the difference between the −2LLs of the final model and each reduced model.

[b]The variability explained was calculated by taking the percentage of the χ^2 value of each measure over the sum of all χ^2 values for all measures ($\Sigma\chi^2$).

might to some degree be influenced by opening and closing gestures, respectively, and thus carry two different types of information: the opening gestures might be more prosodically controlled as they may contain more information that is functional in linguistic terms, while the closing gestures might contain more speaker-specific information. According to the motor program theory, the central nervous system of the speaker actively plans and controls the articulatory behaviors in order to reach articulatory targets.[14,15] It seems plausible that such targets co-occur with mouth opening turning points which again co-occur with vocalic intensity peaks in the acoustic signal. To maximize mutual intelligibility, speakers of the same language should behave more similarly while reaching the same target. Once the target has been reached, the speaker may reduce the degree of control over the articulators, thereby producing movements which are determined more by the ontogenetic biophysical properties (e.g., the mass, damping, and friction) of their bones and muscles. In other words, these two processes are possibly influenced by two properties of the motor plant: controllable properties and intrinsic properties.[11] We argue that the controllable properties play a larger role in the opening gestures, while the intrinsic properties play a larger role in the closing gestures.

Our findings may be of particular interest to research where the identity information about a speaker matters, such as forensic phonetics and automatic speaker recognition. Our results showed that negative dynamics reveal more between-speaker variability than positive dynamics. This means that different parts of the signal intensity contour are more suitable for obtaining speaker-specific information. As such, these parts of the contour might be particularly relevant for forensic speaker comparisons or automatic speaker recognition. A related approach has been shown by Adami *et al.*[12] who fitted a single regression line over the entire energy contour of each syllable to model speaker individuality. The model may perform even better if features pertaining to negative dynamics were included. The theoretical implications of our findings, in particular the assumed relationship between articulatory movements and intensity dynamics requires further-in-depth research:

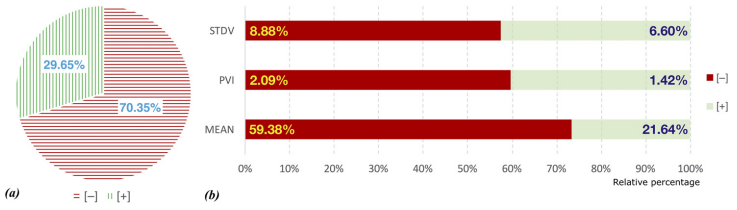


Fig. 3. (Color online) (a) Pie chart showing the amount of between-speaker variability explained by measures of positive dynamics ([+], vertical lines) and negative dynamics ([–], horizontal lines), respectively. (b) Stacked bar chart illustrating relative contributions of both dynamics within the same types of measures; absolute contributions are shown in numbers in each bar.

- We need to take into consideration that there are a variety of factors contributing to the variability of intensity levels in speech. Apart from the size of mouth aperture, there are factors like vocal effort, inherent vowel intensity, prosodic stress and accent or phonotactic arrangements of consonant-vowel sequences. It is imperative that we learn more about the complex relationships between these factors and the actual role that individual movements leading to mouth aperture size play in the individuality of intensity contour characteristics.
- It will be essential to examine the relationships between intensity dynamics and articulatory behavior with articulatory measurement procedures in which the effects of the trajectories of a variety of articulators on the intensity contours are tested. It will also be crucial to learn from such articulatory measurements to what degree the possible articulatory movements contributing to intensity contour variability are intrinsic and to what degree they are acquired behaviors.
- To generalize our findings, replications of results with languages other than our test language (Zürich German) is necessary. Such languages should ideally have different phonological complexities like vowel reductions, consonantal cluster complexities or word stress or accent variability that all might have an impact on articulatory movements and intensity contours.
- So far, we have studied rehearsed read speech only. It seems plausible that articulatory movements are more tensely controlled when the speech needs to be planned during the production process like in spontaneous speech. This speech is also characterized by hesitations, false starts and filled pauses which might have a strong influence on articulatory control.[18]
- Further research is needed to examine, for example, how intensity contours are affected by different forms of signal distortions, especially distortion that can directly affect amplitude envelopes non-linearly, such as dynamic range compressions.

Acknowledgments

This research was supported by the Gebert Rüf Stiftung (No. GRS-027/13) and the Swiss National Science Foundation (No. 100015_135287). We wish to thank Richard Rhodes for helpful suggestions on earlier drafts of the paper, and Adrian Leemann and Marie-José Kolly for their work in constructing the corpus.

References and links

[1]V. Dellwo, M. Huckvale, and M. Ashby, "How is individuality expressed in voice? An introduction to speech production and description for speaker classification," in *Speaker Classification I: Fundamentals, Features and Methods*, edited by C. Müller (Springer, Berlin, Germany, 2007), pp. 1–20.

[2]C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, and A. A. Ghazanfar, "The natural statistics of audiovisual speech," PLoS Comput. Biol. **5**, e1000436 (2009).

[3]A. Eriksson, "Aural/acoustic vs. automatic methods in forensic phonetic case work," in *Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism*, edited by A. Neustein and H. A. Patil (Springer, New York, 2012), pp. 41–69.

[4]T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," Speech Commun. **52**, 12–40 (2010).

[5]A. Leemann, M.-J. Kolly, and V. Dellwo, "Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison," Forensic Sci. Int. **238**, 59–67 (2014).

[6]V. Dellwo, A. Leemann, and M.-J. Kolly, "Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors," J. Acoust. Soc. Am. **137**, 1513–1528 (2015).

[7]L. He and V. Dellwo, "The role of syllable intensity in between-speaker rhythmic variability," Int. J. Speech Language Law **23**, 243–273 (2016).

[8]R. Kanai and G. Rees, "The structural basis of inter-individual differences in human behavior and cognition," Nat. Rev. Neurosci. **12**, 231–242 (2011).

[9]D. A. Winter, *Biomechanics* and *Motor Control of Human* Movement, 4th ed. (John Wiley and Sons, Hoboken, NJ, 2009), 320 pp.

[10]P. Perrier and R. Winkler, "Biomechanics of the orofacial motor system: Influence of speaker-specific characteristics on speech production," in *Individual Differences in Speech Production and Perception*, edited by S. Fuchs, D. Pape, C. Petrone, and P. Perrier (Peter Lang, Frankfurt, Germany, 2015), pp. 223–254.

[11]P. Perrier, "Gesture planning integrating knowledge of the motor plant's dynamics: A literature review for motor control and speech motor control," in *Speech Planning and Dynamics*, edited by S. Fuchs, M. Weirich, D. Pape, and P. Perrier (Peter Lang, Frankfurt, Germany, 2012), pp. 191–238.

[12]A. Adami, R. Mihaescu, D. Reynolds, and J. J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (Hong Kong, 2003), pp. 788–791.

[13]Q. Summerfield, "Lipreading and audio-visual speech perception," Philos. Trans. R. Soc. London B **335**, 71–78 (1992).

[14]P. Birkholz, B. J. Kröger, and C. Neuschaefer-Rube, "Model-based reproduction of articulatory trajectories for consonant-vowel sequences," IEEE Trans. Audio Speech Language Processing **19**, 1422–1433 (2011).
[15]C. Ghez and J. Krakauer, "The organization of movement," in *Principles of Neural Science*, 4th ed., edited by E. R. Kandel, J. H. Schwartz, and T. M. Jessell (McGraw-Hill, New York, 2000), pp. 654–673.
[16]C. C. Creel and M. A. Tumlin, "On-line acoustic and semantic interpretation of talker information," J. Mem. Language **65**, 264–285 (2011).
[17]E. Grabe and E. L. Low, "Durational variability in speech and rhythm class hypothesis," in *Laboratory Phonology*, edited by C. Gussenhoven and N. Warner (Mouton de Gruyter, Berlin, Germany, 2002), Vol. 7, pp. 514–546.
[18]G. P. M. Laan, "The contribution of intonation, segmental durations and spectral features to the perception of a spontaneous and a read speaking style," Speech Commun. **22**, 43–65 (1997).

Appendix I – Terms related to acoustic intensity analysis

Appendix I serves to show the relationships among some closely related physical and psychophysical concepts, including force, pressure, work, energy, power, amplitude, magnitude, intensity, decibel, and loudness (including phon and sone measures). The following texts were extensively referred to when creating this appendix: Huckvale [1], Speaks [2], and Moore [3]. Units in both MKS (meter-kilogram-second) and CGS (centimeter-gram-second) systems are shown.

1. Force

Any object, may it be a celestial body, or an air molecule, stays in rest or uniform motion without force acting upon it. A force applying to such an object causes it to accelerate or decelerate. The size of acceleration or deceleration is proportional to the force applied, but inversely proportional to the mass of the object. The MKS unit of force is Newton. 1 Newton [Force] = 1 kilogram [Mass] × 1 meter/second^2 [Acceleration]. Alternatively, the CGS unit of force is dyne. 1 dyne = 1 gram × 1 centimeter/second^2. 1 dyne = 10^{-5} Newton.

2. Pressure

Pressure refers to the amount of force applied to a unit area. The MKS unit of pressure is Pascal. 1 Pascal [Pressure] = 1 Newton [Force] / meter^2 [Area]. The CGS unit of pressure is dynes/centimeter^2, 1 dyne/cm^2 = 0.1 Newton/meters^2 = 0.1 Pascal.

3. Work

When a force enables a movement of the object that the force acts upon, work is accomplished. The MKS unit of work is Joule. 1 Joule [Work] = 1 Newton [Force] × 1 meter [Displacement]. The CGS unit of work is erg. 1 erg = 1 dyne × 1 centimeter. 1 erg = 10^{-7} Joule.

4. Energy

Energy refers to something that can produce a change, such as a displacement of an object. When such a change occurs, work has been done. Energy is a measure of the capability to do work. The unit of work is also Joule and erg.

5. Power

Power refers to the rate of energy consumption or production. The MKS unit of power is Watt. 1 Watt [Power] = 1 Joule [Energy] / 1 second [Time]. The CGS unit of power is ergs/second, 1 erg/second = 10^{-7} Watt.

6. Amplitude

Amplitude refers to the size of the variation in the signal. For a sound pressure signal, amplitude is often measured in Pascals. When the pressure signal is transduced in electronic domain, it is measured in Volts. Amplitude is a vector, meaning that it can be either positive or negative.

7. Magnitude

Magnitude is the absolute value of amplitude.

8. Intensity

Intensity refers to the amount of power applied per unit area. It is measured as Joule × second^{-1} × meter^{-2}, which is equivalent to Watt × meter^2 in MKS. Expressed in CGS unit, 1 Watt/meter^2 = 1,000 ergs × second^{-1} × centimeter^2.

9. Intensity level (in decibels)

In practice, it is common to express intensity in logarithmic scales with 10^{-12} Watt/meter^2 as the reference, which is equivalent to a pressure of 2×10^{-5} Newton/meter^2 (i.e., 20 μPa, the threshold of hearing at 1,000 Hertz), namely the sound pressure level (SPL).

Number of decibels	$= 10 \cdot \log_{10}(I/I_0)$ $= 10 \cdot \log_{10}(P/P_0)^2$	I -- measured sound intensity I_0 -- reference intensity (10^{-12} W/m^2) P -- measured sound pressure P_0 -- reference pressure (20 µPa)

Another less frequently used unit is the neper. The number of nepers = ln (P/P_0).

In this work, "intensity" is used to mean the intensity level as measured above with the unit of dB re 20 µPa (i.e., dB SPL). Since the microphones were not calibrated for the recordings used in this work, the measures cannot be taken as strict measures of SPL.

10. Phon

The phon is a unit of the perceptual loudness level. The number of phons of a test sound is the sound pressure level (dB SPL) of a sound at a frequency of 1 kHz (reference sound) that sounds just as loud. Phons are used to create the equal-loudness contours that map the relationships between frequencies and sound pressure levels.

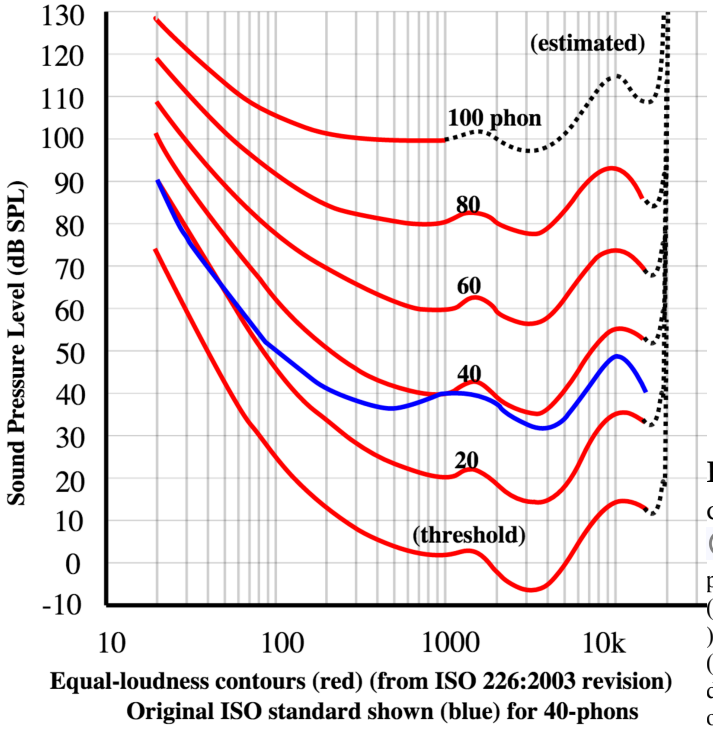


Figure AI-1: Equal-loudness contours

11. Sone

The sone is another unit of perceived loudness. One sone is defined arbitrarily as the loudness of a 1,000 Hz tone at 40 dB SPL. Doubling the perceived loudness results in doubled sone measure.

References

[1] Huckvale, M. (2005). *An Introduction to Acoustics*, 5th edition, London: University College London.

[2] Speaks, C. E. (1992). *Introduction to Sound: Acoustics for the Hearing and Speech Sciences.* San Diego, CA: Springer.

[3] Moore, B. J. (2013). *An Introduction to the Psychology of Hearing*, 6th edition, Leiden and Boston: Brill.

■

Appendix II – Signal processing steps that Praat applies to create an "Intensity" object

This appendix shows in detail the underlying signal processing steps that Praat [1] applies to create the "Intensity" object with default settings as shown in the following screenshot:

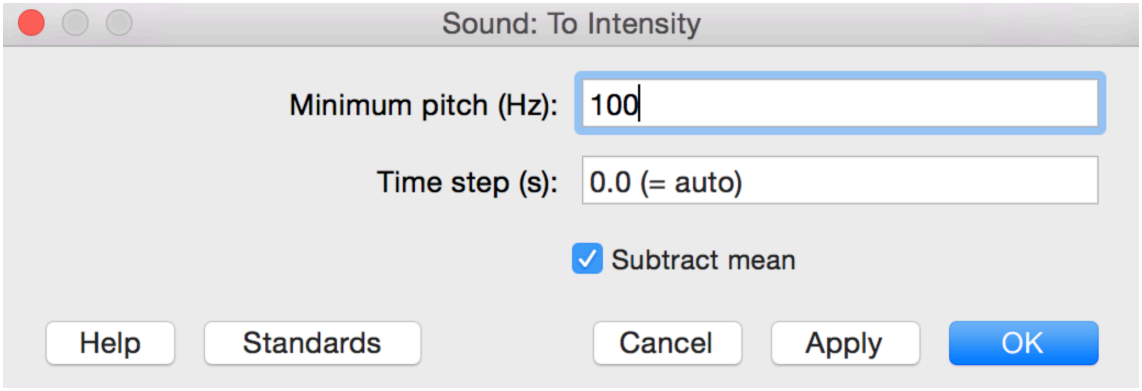


Figure AII-1: Screenshot showing the default settings to create the Intensity object in Praat

1. Removal of DC offset

This step corrects the error induced by the recording equipment that the average amplitude of the signal is not around zero. To remove the offset, the amplitude of each sample of the signal is subtracted by the mean amplitude of all samples of the signal:

$$x[n] = x_{\text{Raw}}[n] - \frac{\sum_{n=1}^{N} x_{\text{Raw}}[n]}{N}$$

$x_{\text{Raw}}[n]$ – Unprocessed signal
$x[n]$ – DC removed signal
n – the index of sample number
N – total number of samples

This step is accomplished by selecting "Subtract mean" in Figure AII-1.

2. Squaring the signal

The amplitude of each sample is squared: $x^2[n]$.

3. Windowing the squared signal

The Kaiser window with coefficient $\beta = 20$ is used to window the signal $x^2[n]$:

$$x_w[n] = w[n] \cdot x^2[n]$$

$x_w[n]$ – the windowed portion of the squared signal $x^2[n]$
$w[n]$ – the Kaiser window function with $\beta = 20$
n – the index of sample number in the windowed frame

The generic function of the Kaiser window $w(n)$ takes the following form [2]:

$$w(n) = \frac{I_0\left(\beta\sqrt{1-\left(\frac{n-N/2}{N/2}\right)^2}\right)}{I_0(\beta)}, n \in [0, N]$$

n – the index of sample number
N – total number of samples within the window
$I_0(\cdot)$ – the zeroth-order modified Bessel function of the first kind
β – the window coefficient

The coefficient β is estimated using the piecewise function:

$$\beta = \begin{cases} 0.1102(\alpha - 8.7), & \text{if } \alpha \in (50, +\infty) \\ 0.5842(\alpha - 21)^{0.4} + 0.07886(\alpha - 21), & \text{if } \alpha \in [21, 50] \\ 0, & \text{if } \alpha \in (-\infty, 21) \end{cases}$$

α – the FIR filter sidelobe attenuation in dB. Praat assumes a sidelobe attenuation of 190 dB [3], therefore β is evaluated as 20.

The idea of using the Kaiser window ($\beta = 20$) is to approximate a Gaussian window. A true Gaussian window has infinite length, which is impossible in applications. The approximated Gaussian window with the Kaiser window ($\beta = 20$) dramatically attenuates the sidelobes flanking the peak in the spectrum, which is desirable in speech signal analysis. Figure AII-2 shows the shape of the Kaiser window ($\beta = 20$) both in time and frequency domains in comparison with other less ideal windows often applied in speech science. Figure AII-3 shows both the waveforms and spectra of a 440 Hz sinusoid, either raw or windowed. It can be appreciated that using the Kaiser window ($\beta = 20$) yields the best result.

The window function that Praat applies to create the Intensity object remains opaque: no window settings are available in Figure AII-1. The user has to write an ad hoc script to create an intensity object using window functions other than the Kaiser ($\beta = 20$). However, the user can specify the window length and the amount of overlap between windows.

The window length is determined by the "Minimum pitch" setting in Figure AII-1. The default minimum pitch is 100 Hz, which allows an effective window length of $3.2 \div 100 = 0.032$ s = 32 ms [3, 4]. The amount of overlap between windows is determined by the "Time step" setting in Figure AII-1. The default "0.0" assumes a

window forward of a quarter of the effective window length, namely 32 × ¼ = 8 ms. Therefore, the amount of overlap between windows is (32-8) ÷ 32 = 75% [3, 4].

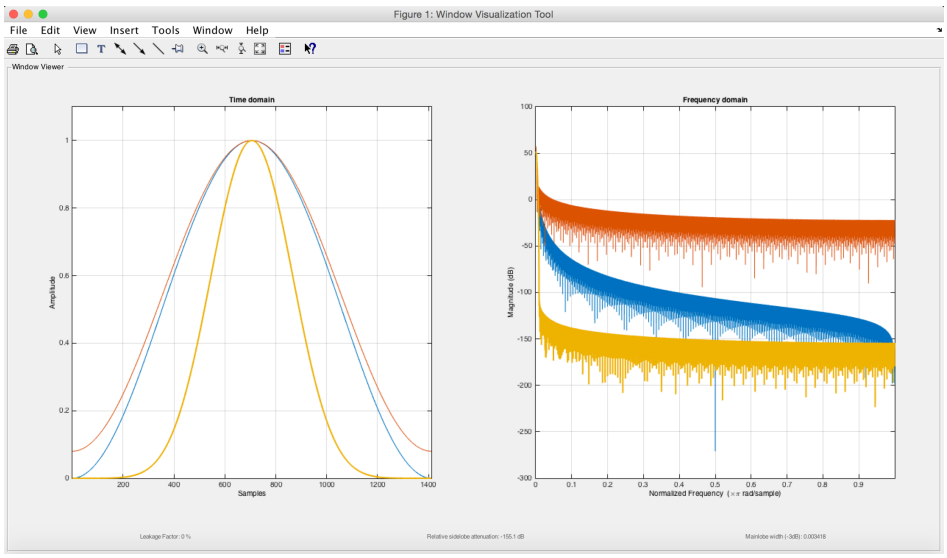
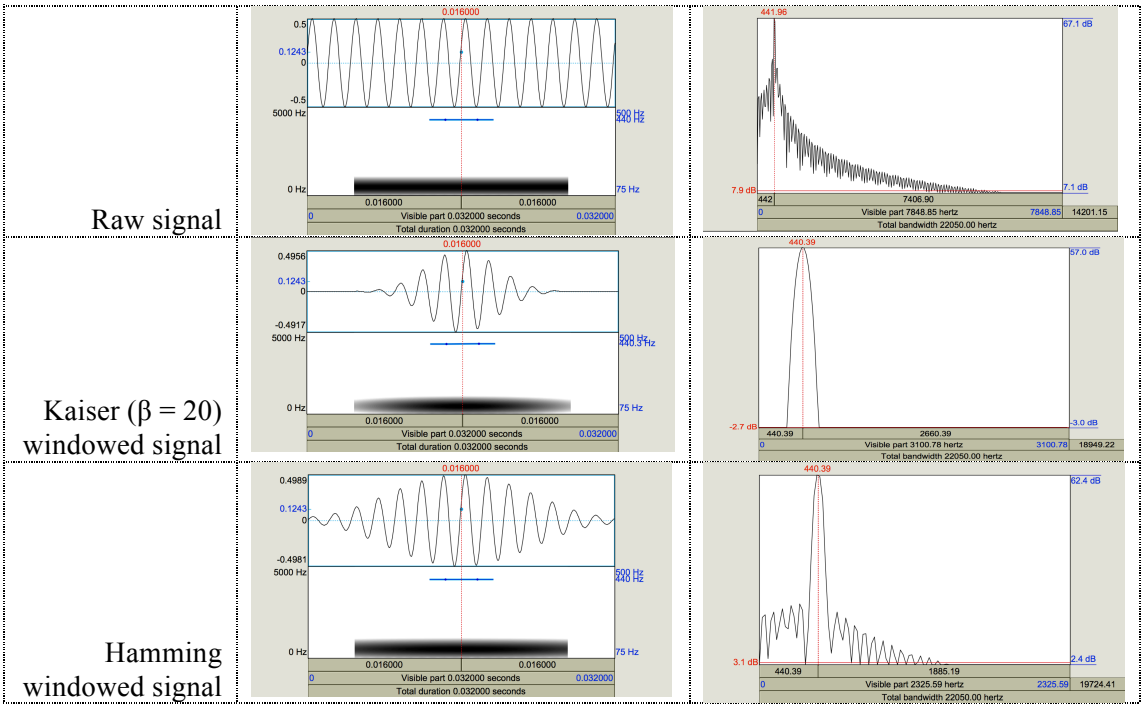


Figure AII-2: Shapes of the Kaiser (β = 20), Hamming and Hanning windows in both time and frequency domains.

Figure created using MATLAB® with the following codes:

```
>> Fs = 44100;
>> w1 = kaiser(round(Fs*0.032), 20);
>> w2 = hamming(round(Fs*0.032));
>> w3 = hanning(round(Fs*0.032));
>> wvtool(w3,w2,w1)
```



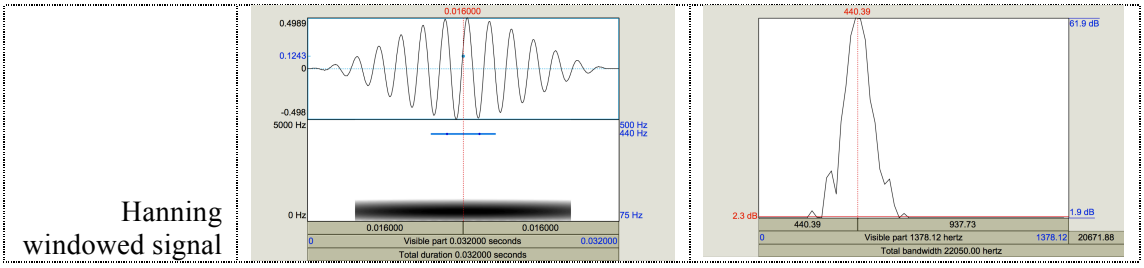


Figure AII-3: The waveforms and spectra of a 440 Hz pure tone (32 ms), either unwindowed or windowed by the Kaiser (β = 20), Hamming and Hanning functions.

4. Creating the Intensity object

Within each windowed frame, the sum of all sample values is taken:

$$SS[k] = \sum_{n=1}^{N_k} x_w[n]_k$$

$SS[k]$ – the sum of all sample values in the k^{th} windowed frame
n – the index of sample number in the k^{th} windowed frame
N_k – the total number of samples in the k^{th} windowed frame
$x_w[n]_k$ – the k^{th} windowed frame

Please note that the signal has been squared before windowing; therefore, the $SS[k]$ basically indicates sum of squares. Finally, the Intensity object is obtained as [4]:

$$Ints[k] = 10\log_{10}\{\frac{SS[k]}{(2\times 10^{-5})^2}\times\frac{1}{d}\}$$

$Ints[k]$ – the Intensity object created by Praat
k – the index of windowed frames
$SS[k]$ – the sum of sample values in the k^{th} frame
d – the window length

References

[1] Boersma, P., and Weenink, D. (2014). Praat: doing phonetics by computer (version 5.3.65), Retrieved from http://www.praat.org/.

[2] MATLAB® online documentation by MathWorks®: "Kaiser window". URL: http://ch.mathworks.com/help/signal/ref/kaiser.html?refresh=true Accessed 30 December 2015.

[3] Praat online manual: "Sound: To Intensity…". URL: http://www.fon.hum.uva.nl/praat/manual/Sound__To_Intensity___.html

Accessed 30 December 2015.

[4] Weenink, D. (2015). *Speech Signal Processing with Praat*, version 18 September 2015, URL: http://www.fon.hum.uva.nl/david/sspbook/sspbook.pdf Accessed 30 December 2015.

■

Curriculum Vitæ

HE Lei 何磊
Born 01.06.1984 in China
Chinese citizen

ORCID: http://orcid.org/0000-0002-9552-9075
ResearcherID: http://www.researcherid.com/rid/H-9506-2017

Education

2012 – 2016
Doctoral student in General Linguistics, Doktoratsprogramm Linguistik, University of Zurich, Switzerland.

2009 – 2010
MSc in Developmental Linguistics, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, United Kingdom.

2007 – 2009
MA in Linguistics and Applied Linguistics, School of Foreign Languages, Renmin University of China, Beijing.

2003 – 2007
BA in English Language and Literature, College of Foreign Languages, Inner Mongolia University, Hohhot, China.

Memberships

Acoustical Society of America (ASA) – since 2012

International Speech and Communication Association (ISCA) – since 2014

International Association for Forensic Phonetics and Acoustics (IAFPA) – since 2014

Academic services

Reviewer for the student poster competition (speech communication sessions) of the 173rd Meeting of the Acoustical Society of America (ASA, Boston, June 25-29, 2017).

Reviewer for INTERSPEECH 2017 (Stockholm)

Publications

• Journal articles

He, Lei ▪ Dellwo, Volker (2017) Between-speaker variability in temporal organizations of intensity contours. Journal of the Acoustical Society of America 141(5), pp. EL488-EL494.
DOI: 10.1121/1.4983398

He, Lei ▪ Dellwo, Volker (2016) The role of syllable intensity in between-speaker rhythmic variability. International Journal of Speech, Language and the Law 23(2), pp. 243-273.
DOI: 10.1558/ijsll.v23i2.30345

He, Lei (2011) Metacognition in EFL pronunciation learning among Chinese tertiary learners. Applied Language Learning 21(1-2), pp. 1-27.
DOI: 10.5167/uzh-128569

• Proceedings and book chapters

He, Lei ▪ Dellwo, Volker (2017) Amplitude envelope kinematics of speech signal: parameter extraction and applications. In Jürgen Trouvain, Ingmar Steiner & Bernd Möbius (Eds.) Elektronische Sprachsignalverarbeitung 2017 (Studientexte zur Sprachkommunikation - Band 86). Dresden, Germany: TUDpress, pp. 107-113.
DOI: 10.5167/uzh-136290

Pellegrino, Elisa ▪ He, Lei ▪ Dellwo, Volker (2017) Computation of L2 speech rhythm based on duration and fundamental frequency. In Jürgen Trouvain, Ingmar Steiner & Bernd Möbius (Eds.) Elektronische Sprachsignalverarbeitung 2017 (Studientexte zur Sprachkommunikation - Band 86). Dresden, Germany: TUDpress, pp. 246-253.
DOI: 10.5167/uzh-136466

He, Lei ▪ Dellwo, Volker (2016) A Praat-based algorithm to extract the amplitude envelope and temporal fine structure using the Hilbert transform. In Proceedings of INTERSPEECH 2016 (pp. 530-534), San Francisco, USA.
DOI: 10.21437/Interspeech.2016-1447

He, Lei ▪ Glavitsch, Ulrike ▪ Dellwo, Volker (2015) Comparisons of speaker recognition strengths using suprasegmental duration and intensity variability: an artificial neural networks approach. In Proceedings of the 18th International Congress of Phonetic Sciences (Paper ICPHS0395), Glasgow, UK.
DOI: 10.5167/uzh-127262

Glavitsch, Ulrike ▪ He, Lei ▪ Dellwo, Volker (2015) Stable and unstable intervals as a basic segmentation procedure of the speech signal. In Proceedings of INTERSPEECH 2015 (pp. 31-35), Dresden, Germany.
DOI: 10.5167/uzh-128615

He, Lei ▪ Dellwo, Volker (2014) Speaker idiosyncratic variability of intensity across syllables. In Proceedings of INTERSPEECH 2014 (pp. 233-237), Singapore.
DOI: 10.5167/uzh-103024

He, Lei (2014) The inadequacy of rhythm metrics to quantify L2 suprasegmental characteristics. In Proceedings of Speech Prosody 2014 (pp. 1095-1099), Dublin, Ireland.
DOI: 10.5167/uzh-128617

He, Lei (2012) Syllabic intensity variations as quantification of speech rhythm: evidence from both L1 and L2. In Proceedings of Speech Prosody 2012 (pp. 466-469), Shanghai, China.
DOI: 10.5167/uzh-127263

Presentations

Pellegrino, Elisa ▪ He, Lei ▪ Giroud, Nathalie ▪ Meyer, Martin ▪ Dellwo, Volker (2017) Speech rhythm and aging *Talk*. Workshop on Speech Perception and Production across the Lifespan (SPPL 2017). London, UK, April 26 - 27, 2017.

He, Lei ▪ Dellwo, Volker (2017) Amplitude envelope kinematics of speech signal: parameter extraction and applications *Poster*. 28. Konferenz Elektronische Sprachsignalverarbeitung (ESSV 2017). Saarbrücken, Germany, March 15 - 17, 2017.

Pellegrino, Elisa ▪ He, Lei ▪ Dellwo, Volker (2017) Computation of L2 speech rhythm based on duration and fundamental frequency *Poster*. 28. Konferenz Elektronische Sprachsignalverarbeitung (ESSV 2017). Saarbrücken, Germany, March 15 - 17, 2017.

He, Lei ▪ Dellwo, Volker (2017) Speaker-specific variability in intensity dynamics *Poster*. The 5th International Winter School on Speech Perception and Production: Learning and Memory. Chorin, Germany, January 9 - 13, 2017.

He, Lei ▪ Dellwo, Volker (2016) A Praat-based algorithm to extract the amplitude envelope and temporal fine structure using the Hilbert transform *Talk*. INTERSPEECH 2016. San Francisco, USA, September 8 - 12, 2016.

He, Lei ▪ Dimos, Kostis ▪ Dellwo, Volker (2016) Between-speaker intensity variability in non-shouted and shouted voice: work in progress *Talk*. The 25th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA). York, UK, July 24 - 27, 2016.

Dimos, Kostis ▪ He, Lei ▪ Dellwo, Volker (2016) Measuring speaker-specific characteristics in shouting with controlled loudness *Talk*. The 25th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA). York, UK, July 24 - 27, 2016.

Milošvić, Milana ▪ Glavitsch, Ulrike ▪ He, Lei ▪ Dellwo, Volker (2016) Segmental features for automatic speaker recognition in a flexible software framework *Talk*. The 25th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA). York, UK, July 24 - 27, 2016.

Pellegrino, Elisa ▪ He, Lei ▪ Dellwo, Volker (2016) A hybrid measure of speech rhythm: possible implications for LADO *Poster*. The 25th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA). York, UK, July 24 - 27, 2016.

Pellegrino, Elisa ▪ He, Lei ▪ Dellwo, Volker (2015) Measuring speech rhythm using both duration and F0: A study on Italian learners of Mandarin *Talk*. Methods in L2 Prosody (ML2P 2015): Romance languages and Chinese at the crossroads. Napels, Italy, November 30 - December 1, 2015.

He, Lei ▪ Glavitsch, Ulrike ▪ Dellwo, Volker (2015) Inter-speaker variability in intensity dynamics *Talk*. The 24th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA). Leiden, The Netherlands, July 8 - 10, 2015.

Dimos, Kostis ▪ He, Lei ▪ Dellwo, Volker ▪ Müller, Mathias (2015) A method for the elicitation of shouted speech with controlled loudness *Talk*. The 24th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA). Leiden, The Netherlands, July 8 - 10, 2015.

Dellwo, Volker ▪ He, Lei ▪ Dimos, Kostis (2015) Listeners can identify speakers based on idiosyncratic rhythm *Talk*. Rhythm Production and Perception Workshop (RPPW 2015). Amsterdam, The Netherlands, July 6 - 8, 2015.

He, Lei ▪ Glavitsch, Ulrike ▪ Dellwo, Volker (2014) Automatic speaker identification using syllable intensity variability: an initial attempt using the kNN classifier *Poster*. Phonetik und Phonologie 10 (P&P 10). Konstanz, Germany, October 9 - 10, 2014.

He, Lei ▪ Dellwo, Volker (2014) Inter-speaker variability of intensity levels across syllables *Poster*. The 23rd Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA), Zürich, Switzerland, August 31 - September 3, 2015.

Dimos, Kostis ▪ He, Lei ▪ Dellwo, Volker (2014) An investigation of the rhythmic acoustic differences between normal and shouted voices *Poster*. The 23rd Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA), Zürich, Switzerland, August 31 - September 3, 2015.

He, Lei (2014) The inadequacy of rhythm metrics to quantify L2 suprasegmental characteristics *Poster*. Speech Prosody 2014, Dublin, Ireland, May 20 - 23, 2014.

He, Lei (2012) Syllabic intensity variations as quantification of speech rhythm: evidence from both L1 and L2 *Poster*. Speech Prosody 2012, Shanghai, China, May 22 - 25, 2012.

He, Lei (2011) Interlanguage rhythm: a durational metrics study among native speakers of Mandarin and Cantonese learning English *Talk*. The 16th World Congress of Applied Linguistics (AILA), Beijing, China, August 23 - 29, 2011.

He, Lei (2010) Metacognition in EFL pronunciation learning among Chinese tertiary learners *Talk*. Languages of Education: The Chinese Context (LECC), Hong Kong, China, October 22 - 23, 2010

■